DATA FOR THE PEOPLE

HOW TO MAKE OUR POST-PRIVACY ECONOMY WORK FOR YOU

大数据和我们

如何更好地从后隐私经济中获益？

[美] 安德雷斯 · 韦思岸（Andreas Weigend）著

胡小锐 李凯平 译

数据是未来的新石油

风靡斯坦福大学的社交数据革命课

亚马逊前首席科学家、大数据专家心血力作

简体中文版全球首发上市



中信出版集团 CHINA CITIC PRESS

VISIT…

LANZAROTE Caliente.com

DATA FOR THE PEOPLE

HOW TO MAKE OUR POST-PRIVACY ECONOMY WORK FOR YOU

大数据和我们

如何更好地从后隐私经济中获益?

[美] 安德雷斯·韦思岸(Andreas Weigend)著

胡小锐 李凯平 译

数据是未来的新石油

风靡斯坦福大学的社交数据革命课

亚马逊前首席科学家、大数据专家心血力作

简体中文版全球首发上市



中信出版集团 CHINACITICPRESS

版权信息

书名：大数据和我们

作者：[美]安德雷斯·韦思岸

ISBN：9787508669694

中信出版集团制作发行

版权所有•侵权必究



当你的一切都被记录在案时

信息本身已成为世界上最大的一桩生意，人们对自己的了解还不如数据银行。数据银行记录的个人数据越多，我们就越缺少存在感。

——马歇尔·麦克卢汉（Marshall McLuhan）

1949年，我父亲还是一个23岁的小伙子，在民主德国当老师。他刚到执教学校所在的小镇时，需要找一个能跟他共住一间宿舍的室友。幸运的是，他在火车站遇到了一个也在寻找住所的人，于是，他们成了室友。但他们搬进住所几天后，父亲就发现他的室友失踪了。这令他十分错愕，之后就开始为他的室友担心。

不久后的一天早晨，父亲正在吃早餐，这时响起了敲门声。也许是室友回来了！父亲打开门，发现来了几个陌生人，他们告诉父亲他获得了教学奖。这是一个十分特殊的奖项，需要亲自授予获奖者本人，他们是来陪我父亲一起去领奖的。父亲对颁奖邀请十分怀疑：当时的场面很奇怪，这些人都阴沉着脸，穿着统一的军用风衣。但是，父亲别无选择，只能跟着他们坐进一辆轿车。上车后，他发现车窗无法从车内打开，他猛然意识到自己被苏联占领军逮捕了。

苏联人指控父亲是美国间谍，证据是他会讲英语。亲朋好友们都不知道父亲身处何地，仿佛他从人间蒸发了。父亲被关进苏联政府管理的一座荒僻的监狱中，在那里遭受了6年的折磨。他根本不知道自己为何被捕，也不知道自己为何获得释放。

透露个人信息会带来杀身之祸，因为数据可以被用来对付我们。实际上，一想起这样的危险，就让我不寒而栗、头脑清醒，因为我知道当时的苏联人是怎样收集数据并用它来对付我父亲的。

在我父亲被关押期间和出狱后，民主德国国家安全部（又称斯塔

西）收集了他的很多信息。两德统一10年后，我请求查阅这些信息。并非只有我一人想知道斯塔西对我的家人所做的一切，柏林墙倒塌后，近300万人都要求查阅本人或其亲属的档案。不幸的是，负责公布斯塔西档案的部门来信告诉我，有关我父亲的所有档案似乎都已被销毁。

但是，这封信中附了一张照片，照片中是斯塔西为我建立的档案。我感到十分诧异，斯塔西竟然有我的档案？我当时只是一个物理学专业的学生。秘密警察早在1979年就开始收集我的信息了，那时候我还是一个懵懂少年。在我搬到美国之后的第二年，也就是1987年，斯塔西更新了我的档案。照片中的我的档案只剩下了一张封皮，我不知道斯塔西收集了我的哪些信息，他们为什么要这么做，要用这些信息做什么。

在斯塔西的势力猖獗之时，收集其“重点关注的公民”的信息是一件非常困难的工作。斯塔西会对这些人跟踪、拍照、截获他们的信件、走访他们的朋友，还会在他们家中安装窃听器，只有这样做才能收集到数据，之后再以纯手工方式对这些数据进行分析。需要收集的数据非常多，在民主德国政府垮台时，全国劳动人口中有1%的人全职服务于秘密警察。但是，这些人还远远不够。德国联邦政府称，民主德国政府在临近垮台之时，有大约20万人为其收集信息。

图0–1 斯塔西为我建立的档案封皮

如今，数据收集人员的工作则简单得多。让我们看几个知名案例。针对美国国家安全局对电话的监听行为，隐私保护活动家经过数月的抗议和法庭交锋后取得了一场小小的胜利。但即便知道自己通话的元数据

会受到美国国家安全局或其他部门的监控，也很少有人会因此退订自己的手机服务。美国加利福尼亚州的一名女销售员提出索赔时称自己因卸载了一款手机应用而遭解雇，因为这款应用无论在上班时间还是闲暇时间都会跟踪定位她的地理位置，并分享给她的主管。新闻曝光脸谱网正在研究人与人之间的情绪影响时，公众愤怒地质问该公司是否在“操控”用户的情感。但是，用户们一如既往地使用脸谱网，而脸谱网也继续在未经用户许可的情况下进行实验，原因很简单，即实验是在线平台设计的重要组成部分。2015年，商业巨头阿里巴巴公司旗下的蚂蚁金服公司在中国推出了一项试点服务——芝麻信用，它通过分析个人交易数据来评估用户的信誉度，这类似于对用户在亚马逊上的购买记录进行评估，以判断该用户是否具备信贷资格。芝麻信用很快在其他领域得以运用，包括在中国一家知名交友网站中作为个人档案的选填项目之一而大受欢迎。没有人会呼吁我们停止使用手机、电子邮箱、导航软件、社交媒体账户、零售网站及其他数字化服务，因为它们让我们的生活更加方便。

我之所以热衷于隐私保护，是因为我发现斯塔西为我建立了档案吗？绝非如此。实际上，斯塔西档案与我每天自愿与他人分享的个人信息相比，不值一提。

从2006年开始，我将自己计划发表的每一场讲座和演讲，以及即将乘坐的每一趟航班信息都发布在我的个人网页上，甚至包括我的航班座位号。我这样做的原因是，我相信通过分享自己的数据所获得的实际价值要高于这样做的风险。数据带来了探索和优化的机会，所以关键问题在于要找到好办法，确保数据使用者的利益与我们的自身利益一致。

我们如何实现这一目标呢？我们可以了解我们分享的是什么数据（和不久的将来可能分享的数据），以及数据公司会如何分析并使用我们的数据。在充分尊重麦克卢汉观点的前提下，我要告诉大家的是，数据公司对我们每个人的数据记录得越多，我们的存在感就越强，我们对自己的了解也越透彻。真正的问题在于，如何确保数据公司对我们的透明性和我们对数据公司的透明性是对等的，还要确保我们对本人数据的使用方式有一定的主导权。本书将告诉我们如何才能实现这两个目标。



社交数据革命
如何确保数据会为我们服务？

> 每一场革命最初都是一个人头脑中的一种思想，一旦同一种思想在另一个人的头脑中出现，它对于这个时代就变得至关重要了。
>
> ——拉尔夫·沃尔多·爱默生（Ralph Waldo Emerson）

早晨6点45分，手机闹钟将我叫醒。于是，我拿起手机，一边浏览电子邮件与脸谱网信息，一边走进厨房，我美好的一天就此开始。手机上的全球定位系统应用软件会记录我的位置变化，并显示出我向东、向北移动了几米。我给自己倒了一杯咖啡，然后走出厨房。这时，手机上的加速计会给出我的行走速度，气压计会记录我何时上楼。由于我在手机上安装了谷歌的应用程序，因此谷歌公司拥有我的这些数据的记录。

吃完早饭后，我要去斯坦福大学上班。在我关灯并拔下移动设备的电源插头后，电力公司安装的“智能”电表就会知道我的用电量开始下降了。当我打开车库门时，电表会探测到与之相匹配的使用签名。当我开车上路时，电力公司已拥有足够的数据断定我已不在家中。当我的手机从另一个基站接收信号时，通信公司也知道我出门了。

驾车行驶在路上时，如果我闯了红灯，安装在街道拐角处的摄像头

就会拍下我的车牌号。谢天谢地，我今天遵纪守法，不会收到交通罚单。但在行驶过程中，我的车牌会多次被拍摄。有些摄像头属于当地政府，有些则属于私营公司，它们通过分析数据了解人们的驾驶习惯，并将此作为产品出售给警方、开发商及其他利益群体。

我到达斯坦福大学时，会使用手机上的“无忧停车”应用支付停车费。停车费自动记入我的银行账户，同时学校的停车管理小组会收到我的付款通知，这样一来，校方与我的开户银行都知道我在上午9点03分到达校园。由于我的手机不再以汽车的行驶速度移动，谷歌公司会推断出我已停车并记录下我的位置，以便我日后查询当时的位置记录。我也可以通过美国车险服务商Metromile公司的保险应用查询我当时所在的位置，这款应用通过我的车载诊断系统实时记录我的驾驶数据。这让我可以立刻发现今天的汽车燃油效率较低——每加仑[注]汽油行驶了19英里[注]，我此次通勤花了2.05美元。

上完课后，我打算和旧金山的新朋友见个面。我们在“虚拟世界”中见过面，当时我们共同的朋友在脸谱网上发了帖子，我们都对它进行了评论，也很赞赏对方的看法。之后，又发现我们在脸谱网上有30多个共同好友，所以我们确实应该见一面。

谷歌地图预计我将在晚上7点12分到达目的地。与往常一样，它的预测误差只有几分钟。这位朋友居住公寓的一层是一家销售烟草产品和吸食大麻器具的商店，而我的智能手机上的全球定位系统应用软件无法区分公寓和商铺。我的车载导航与谷歌导航都告诉我，我今天晚上去了一趟毒品商店——这是我上床前查阅第二天的天气预报时，谷歌广告推送告诉我的。

这不只是一场社交数据革命。

将欲取之，必先予之

每天都有10多亿人像我这样产生和分享社交数据。社交数据是有关你本人的信息，例如你的运动、行为、兴趣，以及你和其他人、地点、产品，甚至意识形态之间的关系。其中有些数据是在你本人知情的前提下自愿分享的，例如在使用谷歌地图时登录并键入目的地；其他数据则

并非如此，你经常会在不经意间就分享了自己的数据，这是享受互联网与移动设备所带来的便捷性过程的重要部分。显然，在某些情况下，分享数据是你获取服务的必要条件：如果你不向应用软件提供你当前所在的位置和目的地，谷歌公司就无法为你找出最佳的行车路线。在某些情况下，你可能很乐意提供信息，例如你给某个朋友在脸谱网上的发帖点赞或在领英网上对同事的工作表示肯定，以表明你愿意以某种方式鼓励和支持他。

社交数据有时可以做到比较精准，能将你的位置精确到1米之内。但是，在通常情况下，社交数据都很粗略，有时也不够完整。例如，除非我登录可以显示家中智能电表读数的某个应用（比如，为了查看我在去机场之前是否将家中所有的灯都关上了），电力公司才能知道我何时离家，但也仅限于此。这种数据过于粗略，也许对我没有太大的帮助。与此相似，我在拜访旧金山的那位新朋友时，虽然社交数据可以准确地显示出我所在位置的经度和纬度，但对我当晚活动的推测却是完全错误的。有时候，虽然数据看似十分精确，但在很大程度上这是数据解读的结果。实际上，社交数据本身是非常粗略的。粗略的数据很可能不完整、易出错，有时其中还会掺杂欺诈数据。

无论是被动还是主动分享的数据、强制还是自愿分享的数据、精确还是粗略的数据，社交数据的总量呈指数增长趋势。如今，社交数据总量翻一番所需的时间只有18个月。在未来5年内，社交数据总量将增长约10倍，或者说增长一个数量级；在未来10年内，社交数据总量将增长约100倍。换言之，2000年全年产生的数据总量目前只需要1天即可完成。以这样的增长速度计算，预计到2020年，不到1个小时就能产生等量的数据。

要知道，“社交数据”并非仅适用于社交媒体的流行词汇，这一点很关键。许多社交媒体平台的设计旨在进行播报，以推特为例，沟通几乎总是单向进行的，由名人、权威人士或营销人士向公众传播信息。社交数据更加民主化，你可以通过推特或脸谱网分享你的信息、所在公司的信息、你的成果、你的看法，但你的电子踪迹比这些更深远。根据你在谷歌网站上的搜索记录、你在亚马逊网站上的购买记录、你在讯佳普（Skype）上的通话记录、你手机的实时定位，再将这些信息与其他多种渠道相结合，就能得出有关某个人的一幅独特的“肖像画”。

此外，社交数据不会止于你本人。在你展示自己通过与亲朋好友、

工作同事的沟通建立起的亲密关系时，你便创建并分享了数据。你所创建的社交数据不仅涉及友人，也会涉及陌生人，例如你在评价某件商品或在照片墙（Instagram）上传照片时。空中食宿（Airbnb）是一个租用房间或套房的应用平台，你若要注册账户就需要验证身份——不仅要使用政府核发的身份证，还要使用你的脸谱网账户。社交数据正在嵌入你家中的智能温度计、汽车的导航系统以及职场的办公软件，并开始成为教室与医院诊疗室中的亮点。随着手机配备了越来越多的传感器和应用，它们可在我们的家中、商场或单位里跟踪我们的一举一动。你将越来越难以掌控有关你日常活动的数据，甚至包括你内心中最隐秘的愿望。数据科学家将化身为侦探与艺术家，通过人们留下的电子踪迹为他们绘制出越发清晰的行为素描画。

通过检查并提炼这些电子踪迹，可以发现人们的偏好或倾向，还能做出预测，例如人们可能会购买何种商品。在我担任亚马逊公司首席科学家期间，我与杰夫·贝索斯共同制定了该公司的数据战略和以客户为中心的文化。我们开展了一系列实验，比较网站编辑或消费者所写的商品评论中哪一种会让客户更开心，并观察依据传统的人口统计信息或个人点击情况为客户做推荐是否成功率更高。在举办厂商赞助的促销活动时，我们发现真正的沟通可以爆发出巨大的力量。我们为亚马逊开发个性化工具，使人们做出购买决定的过程及所购买的商品都产生了根本性改变，并且成为电子商务的标准。

离开亚马逊之后，我在斯坦福大学和加利福尼亚大学伯克利分校为成千上万的本科生和研究生开设了社交数据革命课程，还在中国上海的复旦大学与中欧国际商学院、北京的清华大学教授这门课程。我同时继续经营社交数据实验室，成员是我在2011年结识的一群数据科研人员与思想领袖。在过去10年里，与我合作的公司包括阿里巴巴、美国电话电报公司、沃尔玛、美国联合健康保险集团，以及一些大型航空公司、金融服务公司、交友网站。我积极倡导把数据的决策权与客户或用户分享，他们是与你我一样的普通人。

没有人能够独自处理当下的所有数据并做出明智的决定。但在让数据服务于我们的需要和解决问题的过程中，谁能够获得必要的工具呢？从这些数据中分析得出人们的偏好、倾向和做出预测后，是将其提供给少数强大的组织，还是提供给所有人使用呢？使用社交数据所需支付的费用是多少呢？

随着我们逐渐认识到社交数据的价值，我相信我们的重点不仅是获取数据，还必须采取某些行动。我们每天都会做出很多决定，而有些决定一生中只会做一次。但是，这并不意味着今天产生的社交数据的寿命很短。我们今天的行为方式可能会影响我们今后几十年的选择，很少有人能在短期或长期内观察到自己的所有行为或分析出这些行为将如何影响自己。社交数据分析有助于我们找出各种可能性，但必须经过深思熟虑方可做出最终选择。

毕竟，这些科技无法了解我们每个人乃至整个社会对未来生活的憧憬。许多国家都出台了法律，保护个人在就业或医疗方面不受歧视。未来某一天，这些法律或许将不复存在（在某些国家，直到现在也没有这样的法律）。假设你希望获得有关减肥和锻炼的建议，于是你决定在医疗应用或网站上表达自己对胆固醇过高的担心。这样做会不会对你不利呢？如果法律规定，在医生向你告知健康风险并推荐健康的生活方式之后，你仍然不愿意放弃吃油炸食品，依旧喜欢瘫坐在沙发上，就可以依法对你收取更高的医疗费用，你怎么办？如果你的主管利用某种服务软件在网上查找有关你的信息，他可能认定你的生活方式不适合在他的公司任职，从而拒绝考虑你的求职申请，你怎么办？这些都是实实在在的风险。

如果这些数据是你独立创建并透露出去的，那么，一旦察觉到风险，你或许可以停止这种行为。这会给你带来许多不便，却是可行的。但是，人们对有关自己的许多数据并没有掌控力。由于社交数据被公司和政府用于改善结果、提高效率，因此我们更不可能掌控这些数据。

社交数据关乎社会大众，我们每个人都需要考虑怎样做才是最好的数据利用方式。科技正在飞速发展，收集和分析数据的公司主要从事信息的产出与编码，并不负责制定原则。即使它们考虑那些原则性问题，也仅仅是因为业务需要而临时为之。对人类未来会产生重大影响的原则性问题的决定权，绝不应该交到数据公司手中。

我们可以允许对所有这些数据进行收集、合并、汇聚、分析，以便能在决策过程中更好地做出取舍。取舍是任何重要决策的必要组成部分，在做取舍时，人的判断十分关键。我们的生活不应由数据来驱动，而应让数据为我们的生活服务。

后隐私时代的原则

我们已经认识到数据在生活中发挥着越来越重要的作用，也已经采取了许多措施保护自身的利益。20世纪70年代，美国与欧洲针对信息的公平使用采取了大体相似的原则。人们有权知道谁在收集自己的数据以及这些数据的使用情况，当发现数据不准确时，还可以要求修正数据。然而，对于今天的新型数据来源与分析方法，这些保护措施要么过于严厉，要么过于无力。

之所以说它们过于严厉，是因为这些措施都想当然地认为可以对收集到的所有数据添加标签。亚马逊公司可能会以浅显易懂的术语，准确地解释它是如何使用收集到的个人信息的，它甚至能用这些信息帮助人们做出更明智的决定。但是，对这些信息进行审查需要大量的时间。我们中有多少人会花时间对所有的相关数据进行核查呢？查阅亚马逊公司怎样对每个数据点分配权重，会给你带来什么好处，还是说你宁愿亚马逊交给你一份数据使用简报呢？

之所以说这些保护措施过于无力，是因为即便你能够核查你创建和分享的所有数据，你也无法全盘掌握你的所有相关数据，因为这其中包括其他人创建和分享的关于你的数据，包括你的亲朋好友、同事、老板。你在网上访问的公司和你在实体世界中访问的大部分公司也会产生（有时也会分享）关于你的数据。你在街上遇到的陌生人以及和你打交道的其他许多公共组织和私营组织，同样如此。谁来判定这些数据的准确性呢？今天的数据来自诸多层面，人们无法拥有充分的权利来修正关于自己的数据。此外，即便是准确的数据也可能对你不利。

数据的产生、沟通、处理过程中会发生巨大的定量变化和定性变化，仅有知情权与修正数据权显然是不够的。迄今为止，试图修改这些指导原则的努力几乎全都集中于个人掌控权与隐私权这两个方面。不幸的是，其理念与实践从技术上看已经落后达一个世纪之久。而且，控制与隐私权的标准迫使人们与数据公司签下不平等合约。如果你希望用数据改善你的决策过程，你就必须同意按照数据采集者的条款收集自己的数据。一旦你这样做，就说明数据公司已经按照法律规定为你赋予了个人数据的“控制权”，而无须考虑你是否真正拥有选择权或你的个人隐私权是否会受到影响。如果你希望保护个人隐私，就不应该同意数据公司收集你的数据，但这会牺牲你对相关数据产品与服务的使用权，降低你

从自己的数据中所能获得的价值。只有这样，你才能对你的数据保持掌控权。

如今，我们需要做的是制定一套标准，帮助我们评估因分享和收集数据所产生的风险与回报，同时拥有对数据公司进行问责的权利。基于20年来与数据公司的合作经验，我认为透明性与主动性原则最有可能保护我们免遭社交数据滥用的伤害，并能提高我们从中得到的价值。

透明性涵盖了个人对自己数据的知情权：内容是什么？用途是什么？对用户的好处是什么？数据公司是躲在单向镜的另一面暗中窥探你的隐私，还是也给你打开一扇窗户，让你看到它们如何使用你的数据，从而判断该公司的利益是否（以及何时）与你的自身利益一致？你需要分享多少数据，方能获得你想要的数据产品或数据服务呢？从历史上看，机构与个人之间存在巨大的信息不对称的情况，这使机构占据巨大的优势。机构不仅有强大的能力收集个人数据，还能将你的数据与他人的数据做比较。你需要了解你提供的数据与你得到的数据产品与服务是否对等。

相对于客户与零售商之间的传统关系，亚马逊是如何让购物体验具备透明性的呢？当你打算购买一件商品时，零售商会提醒你之前购买过这件商品吗？这样做会让他损失一单生意。在亚马逊网站上，如果你点击购买已在这家网站上买过的书籍，就会看到网站的提示：“你确定要买这本书吗？你在2013年12月17日买过这本书。”如果你买过某张音乐专辑中的一首歌曲，之后决定购买该专辑的其他所有歌曲时，亚马逊在“完成购买”环节会自动从这张专辑的价格中减去你之前所购歌曲的金额。亚马逊对人们的购买数据采取这种使用方式，是为了最大限度地减少客户的不满。与此相似，大多数航空公司的常飞旅客计划都会给客户发送信息，提醒其即将过期的里程数，而不是放任其作废。

不幸的是，透明性远未成为通行的标准。以给客户服务中心打电话这种典型的体验为例，拨通电话后，你肯定会听到一番警告：“为保证服务质量，本次通话可能会被录音。”你别无选择，如果你想与客服代表通话，就必须接受这样的条款。就算要录音，为什么只有公司有权获得此次通话的录音呢？如果只有通话的一方拥有获取此次通话录音的权利，那么“为保证服务质量”又意味着什么呢？数据对等原则意味着付款的客户同样可以获得录音。

无论何时，只要我听到客户服务代表说通话可能会被录音时，我就会对他说，我也可能会对此次通话录音，以保证我所获得的服务质量。在大多数情况下，客服代表都会选择与我合作，但偶尔也会直接挂断电话。当然，我也会在不征求客服代表同意的情况下自行对通话进行录音，但我要说明一点，这种做法在某些地方是违法的。如果我没有获得客服代表对我承诺的服务时，我就会利用手头的证据向他的主管投诉。如果这样做仍然无效，我就会将音频文件上传到网上，希望通过音频文件的传播催促该公司迅速解决我的问题。就像康卡斯特公司曾经遇到的情况一样，当时客户想取消服务，但是反复遭到该公司的拒绝，最终客户将通话录音发布在推特上才如愿以偿。

通过这种方式，你不必违反法律就可以摆脱不平等的对待。为了让透明性成为新的默认原则，数据公司需要向公众提供更多的信息而不是更少的信息。

但仅有透明性是不够的，你还需要主动性，主动性包含个人根据自己的数据采取行动的权利。数据公司的“默认”设置一目了然吗？你能够出于种种原因修改你的数据吗？你能随心所欲地使用公司产生的数据吗？你是被诱导（或被迫）从有限的几个选项中做出选择（这些选项几乎都更有利于数据公司）吗？你能修改参数并探索不同的情景，以发现种种可能性吗？主动性是个人根据数据公司所发现的关于他的偏好与行为模式进行选择的权利，包括要求数据公司按照他提出的条件向他提供信息的权利。

在基本层面上，主动性关乎人们有能力创造出有利于自己的数据。亚马逊公司一直坚持原原本本地在网页上呈现客户的评论，无论评论是好是坏，是五星还是一星，是为了获得他人的认同还是为了实现成为图书评论家这一人生理想，都没有关系。亚马逊更看重的是，这些评论与其他想购买图书的客户之间的关系。比如，通过评论发现，虽然客户没有选择退货，但对此次购买的商品不太满意。这些数据有助于客户判断某件推荐商品是不是自己的最佳选择，由此亚马逊公司给了客户更大的主动性。

许多营销人士津津乐道于市场定位、市场细分与转化。我不了解你们的想法，但我不想被定位、细分、转化，也不想被剖析，这些不是主动性的表达方式。我们不能想当然地认为每个公司都会主动遵循透明性与主动性原则。我们还必须超越这些原则：我们需要拥有明确的权利，

这有助于我们表达自己的愿望，将透明性与主动性转化为实实在在的工具。

如果我们能促使数据公司同意提供一系列有意义的权利与工具，就能产生我所说的“关系反转”，即对个人与机构之间的传统关系予以逆转。亚马逊公司决定由客户撰写大部分商品评论，这也属于关系反转，社交数据革命将会提供更多这样的机会。随着人们拥有越来越多的工具去帮助自己做出更好的决定，过去公司常用的市场营销手段的效果将会越来越差。由公司告诉处于弱势地位的客户应该购买何种商品，这个时代已经一去不复返了。而且很快，就会由你来告诉公司应该为你做些什么。在某些地方，人们已经体会到这种变化了。

关系反转是物理学家观察世界的重要方式。关系反转经常与相变联系在一起，后者指的是外部条件导致物质属性发生突变——当加热到沸腾状态时，水就会从液态变为气态。日益增加的数据量对社会所产生的影响就相当于物理系统中热量的增加。在某些条件下——当数据公司遵循透明性与主动性原则时，就会产生关系反转，也就是说，这更有利于个人而非公司或公司的首席市场官。

我们所有人的利益都与此次社交数据革命息息相关。如果你希望从社交数据中获益，就必须分享自己的信息。你从社交数据中获得的价值通常在于你拥有了更强的决策能力，即在促成交易的谈判中、购买产品与服务时、申请贷款的过程中、寻找工作时、获取教育与医疗时、改善你所在社区的硬件时，你可以做出更明智的决定。你在分享数据时所付出的代价与承担的风险不应大于你所获得的收益。数据公司收集的数据及其采取的行动应具有透明性，这一点至关重要。另外，你还需要对数据产品与服务拥有一定的掌控力。否则，人们如何判断自己所获得的收益是否大于付出的代价呢？

新的游戏规则

信息是权力的中心。如果你拥有的信息比别人多，那么你很可能从中获利，这就像二手车销售员将劣质车推销给不明真相的客户一样。随着沟通与处理过程变得越来越便宜和普及，巨大的信息不对称风险发生的可能性将越来越大，因为没有人能够掌握所有数据。

在这些产生和分享的数据中，有许多是关于我们的个人生活的：居住的地点、工作的地点、前往的地点，喜欢的人、不喜欢的人、陪伴的人、共进午餐的人，运动量、服用的药物、家用电器、触动心灵的杂志。我们的生活在数据公司面前是透明的，这些公司收集并分析我们的数据，有时它们还会私自销售我们的数据，或者擅自保存我们的数据。在个人数据被修改、交换、销售的过程中，我们需要拥有一定的发言权，此外我们还要对个人数据的使用制定更多的条款。双方（数据创建者与数据公司）都必须遵循透明性与主动性原则。

这需要我们从根本性上改变对数据和自我的看法。在第1章中，我会介绍数据公司分析数据的几种方式，并以提炼过程做类比，说明公司如何将原始数据转化为产品与服务。我将在第2章中讨论个人及其特点，并论述我们在生活中留下的电子踪迹是如何破坏我们的隐私性幻觉的（搜索、点击、评论、使用与刷卡），并在此过程中产生了新的身份概念。无论我们是否愿意，都发出了自己兴趣的真实信号。在第3章中，我将论述重点从个人转移到人与人之间的关系，以及社交网络怎样展示并影响数字化时代的信任。在第4章中，随着各种传感器（不只是摄像头）的联网，我将介绍人们如何利用精确度越来越高的手段记录我们所在的环境，以及数据公司如何分析这些传感器收集的数据，并推断出人们的位置、情绪状态与兴趣。

在此基础上，我提出了6项权利。为了确保未来我们的数据能够真正为我们服务，我认为这些权利十分关键。其中有两项权利是访问数据的权利与核实数据的权利，它们旨在提高透明性。其余4项权利主要通过主动性原则使人们对自身数据具有更强的掌控力，包括修正数据的权利、对数据进行模糊处理的权利、利用数据开展实验的权利、将数据导出给其他公司的权利。通过对个人数据行使上述权利，就能对我们的购物方式、支付与投资方式、工作方式、生活方式、学习方式、使用公共资源的方式等产生影响。在最后一章中，我将论述如何实现这6项权利。

在这个时代的转折点上，人们正在界定创建数据的人与把数据转化成产品和服务的组织之间的关系。不仅游戏规则正在改变，从性质上看，我们正在玩的这个新游戏还要求我们重新界定客户与零售商、投资者与银行、雇主与雇员、患者与医生、学生与老师、公民与政府之间的关系。此时，我们应当表明立场并真正了解数据的用途，以便能够获得

利益并清楚由此产生的结果。只有这样，我们才能评估我们的利益是否与数据公司的利益一致。对于大多数新科技而言，并非机器决定一切。只要人们使用机器，调整自己的期望，并在此过程中修订社会规范，社交数据革命就会悄然而至。

如果我们迎接这场挑战，数据就有可能由取之于民、归之于民演变为用之于民。让我们一起投身于这场革命吧！

---

1. 1加仑≈3.8升。——编者注
2. 1英里≈1.6千米。——编者注

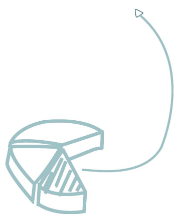


第1章
培养数据素养

数字公民的必备工具

数据公司如何进行数据挖掘？

你的个人数据在它们心中价值几何？



> 在18世纪，如果能够大声朗读《圣经》或者教义问答手册中为人所熟悉的段落，就会被视为有文化。今天，如果有人只会阅读这些内容，却看不懂那些关乎生存的经济信息，就会被归为功能性文盲。
>
> ——乔治·米勒（George Miller）

“让数据为人类造福”不是一句空洞的口号。每天，我们都会面对各种数据产品和服务，收到基于社交数据的各种评分排序和推荐信息。传统意义上的推销“狂人”已经退出历史舞台，取而代之的是数据科学家。他们日复一日地运用各种算法，追踪10亿人口留下的纷纭繁杂的数字痕迹。数据集呈指数增长趋势，但更需要关注的一个重要现象是我们心态上发生的变化。为了全身心地投入社交数据革命，我们必须摒弃被动型“消费者”的陈旧形象。我们不能不加选择地全盘接受，而要变身为积极的社交数据创造者。卖家与买家之间、银行与借贷方之间、公司与员

工之间、医生与患者之间、教师与学生之间，正在通过拉锯式对垒形成势均力敌的状态。这为数据由取之于民、归之于民演变为用之于民创造了条件。

事实上，让数据为人类造福的需求具有无以复加的重要性。作为21世纪最重要的原材料，数据就是新石油。一个多世纪以来，油田的发现与石油开采技术的进步对经济与社会产生了深远的影响。人们通过提取、储存和精炼等环节，把石油变成为人类服务的各种产品。现在，由原始数据转变而成的产品和服务，正在改变我们的生活，其影响力足以与工业革命相媲美。

原油不可以直接利用，我们必须对它进行精炼，将它转变成包括汽油、塑料在内的多种产品。精炼油又为工业时代的机器提供动力，在现代经济活动中发挥了重要作用。同样，原始数据本身并无多大用途。只有经过数据服务商的收集、分析、比较、过滤而产生的数据产品和服务，才具有价值。这些经过处理的数据不能为工业革命时期的机器提供动力，但可以让社交数据革命的大船破浪前进。

幸运的是，数据与石油有几个本质性区别。全世界的石油总量是有限的，剩余的石油资源越少，开采成本就越高。与之相反，数据量呈指数增长趋势，而数据交流与处理所需要的技术成本却在不断下降。至2015年年底，超过50%的成年人拥有智能手机。普通美国人每天使用手机的时间约为2个小时。据估计，我们每天接触电话的次数为200～300次，比大多数人一个月内接触伴侣的次数还要多。我们每次使用电话，都会产生数据。与石油不同的是，数据永远不会枯竭。

石油的使用会受到稀缺性与物质形态的限制；而在使用数据时，我们则需要考虑现代社会的数据是数字形式的，而且数据量极其庞大等问题。一份原油，只有一个实体拥有其使用权，或者只能提炼为一种产品；而一个数据库，多个实体可以同时接入，并且创造出多个不同的产品。我们在制定法律和社会规范时，是以数据供应不足为前提的。例如，由于缺少大量数据，我们创造了保险业，以免一些可怕的事件给我们的生活造成损失和伤害。因为无法确定某个人遭遇入室盗窃或者患糖尿病的可能性，保险公司就把人们集中起来，让他们共担风险，然后按照均价向每个人收取保险费。随着数据越来越多，我们可以做到准确预测每个人面临的风险，做到按人收费。我们可以闭目塞听，假装这些数据不存在；我们也可以承认这些数据确实存在，然后思考我们的人生可

能会因此发生哪些变化。想一想，有了这些新型资源，我们可以创造出一个什么样的世界呢？

如果有合适的工具，新技术可能会增强我们的能力。在古登堡发明印刷机之前，书籍供不应求，广泛传播信息的成本非常高。因此，对绝大多数人而言，花费很长时间去学习阅读是一件徒劳无益的事。在网络发明之前，曾任普林斯顿大学心理学教授的乔治·米勒在其作品中谈到判断一个人是否有文化的现代标准。他为众多毕业生在阅读、数学和科学等方面的素养不够而深感不安，担心他们在一个知识产业占主导地位的经济体里找不到工作。如今，我认为人们同样迫切需要培养一种新素养——数据素养，包括了解数据服务商的工作机制，知道哪些参数可以改变或不可以改变，善于改正错误，了解不确定因素，以及预测分享社交数据可能会带来哪些结果等技能。在当今世界，数据服务商的推荐意见与分析结果对我们的大多数决策都有引导作用，因此，数据素养必不可少。

数据挖掘的力量

亚马逊公司跻身于首批有影响的数据挖掘者行列，这并不奇怪。要在零售业获得成功，就必须清楚应该为潜在顾客提供哪些产品，因此有必要收集库存、价格、广告及顾客购买习惯等数据。

200年前，商家掌握的数据只有两样：货架上的存货和抽屉里的现金。每天打烊时，他用钢笔在纸质分类账簿上记录下这些数据。面对价格相仿的类似产品，顾客只能根据产品承诺是否可靠、包装是否诱人，以及邻居和亲友的口碑做出取舍。大约150年前，一些公司（其中蒙哥马利–沃德公司和西尔斯–罗巴克公司的名头最响）发行了囊括1 000多件产品介绍的邮购目录，这一举措令美国小城镇的顾客非常高兴。这些有创新意识的公司知道每名顾客订购了哪些产品，也知道顾客希望在哪里签收这些产品，因此，它们清楚应该在哪些地方销售哪些产品。100年前，邮购公司开设了实物陈列室和仓库。为了提高产品存储的效益，他们请了一批数据分析师，让他们梳理过往的销售数据，并预测顾客的未来需求。50年前，零售业的情况又一次发生了变化。邮购公司及其临街店铺借助新启用的邮政编码系统，轻松地掌握了美国顾客的消费特征。在随后20年的时间里，公司根据这些行政区划，不断收集顾客的详

细人口统计信息。20世纪60年代中期，信用卡在美国出现，为收集单个顾客的交易数据提供了一个有效的方法。这是网络诞生之前，数据可以提供的最具个性化的信息：居住地、消费金额与消费地点。

创立于1969年的安客诚（Acxiom）与其他数据中间商将家庭或住户调查数据拆分开，然后把每个人归入“苹果派家庭”“蓝血贵族”“快速自提型顾客”“郊区中产阶级妈妈”等几十个顾客群。数据中间商在编制这些标签的时候，只能从一些公开的数据和邮购数据中获取信息。它们按照邮政编码，挨家挨户地查阅资产评估资料，从而了解哪些家庭有游泳池。在消费者数据稀缺的年代，“市场细分营销法”就是上天赐予的神兵利器。截至20世纪末21世纪初，安客诚的年收益增长至10亿美元左右。

数据中间商自然会寻觅一切机会，将市场细分分析的做法延伸至网络零售。在进入亚马逊的前一年，我应邀加入了一个团队，帮助安客诚研究如何在它的邮编数据与基于家庭的数据中添加数码单元。安客诚资方所关心的大问题是，如何为现有客户的家庭购买记录添加正确的邮箱地址。就在安客诚考虑这个细枝末节的问题（在数据库中新添一个数据域）的时候，亚马逊等公司已经准备迈出一大步，跨入社交数据的浩瀚领域了。我清楚地记得，当时（从那以后过了6年时间，第一部苹果手机才面世），我努力地向安客诚管理层解释，在线数据意味着公司能够掌握的信息远不止客户的家庭人口统计资料。零售商将有能力追踪每一个搜索查询、每一次鼠标点击、每一桩交易，以及客户放进“购物车”最后却放弃购买的所有产品。如果公司可以随意获取这些数据，它们就可以真正地向每个顾客推销它们的产品和服务。也就是说，将市场细分成一个个的单一顾客。

备货齐全是亚马逊追求的目标，因此有人把它称为“囊括所有商品的商场”。但是，从更深层次来看，我们也可以把它称为第一家“储存所有数据的商场”，因为亚马逊一直致力于储存顾客与产品的所有数据。亚马逊销售的产品有数千万种，因此它不可能把所有产品都呈现在你的眼前。我们也不可能一页一页地浏览产品目录的全部内容。如果不告诉亚马逊你在寻找什么，它就没有办法为你展示它可以提供并且可能符合你需求的产品。你必须提交数据，才能得到按序排列的搜索结果。在付款之前隐藏自己的购买兴趣，这已经不再是你的可选方案了。

我于2002年进入亚马逊公司，这家公司当时的一个目标是在分析邮政编码的基础上，进一步分析顾客与网站之间达成的每一桩交易。我和

团队成员一起，针对500种个人特点，对每名用户进行鉴别。我们首先会提出若干问题：例如，送货地址与最近的书店（或大商场）之间的距离，对顾客在亚马逊的购物频率或者消费金额是否会产生影响？如果顾客选择用信用卡支付，那么这对她今后的购物行为是否有预测作用？对亚马逊而言，购买过两类或多类产品的顾客，从年销售额的角度看，是否比只购买过书籍的顾客更有价值？每名顾客在订购商品时，白天与晚上做出的选择会不会不同？这些分析结果为公司的众多决策（比如在加大营销投入力度和实施减价策略之间的取舍）提供了信息支持。

此外，有了这些分析结果，我们也可以为顾客推荐合适的商品，帮助他们做出购买决策。我们发现，在预测顾客购买某件产品的可能性时，该顾客的购物记录所发挥的作用往往比不上该产品与其他产品之间的关系。产品之间的关系各不相同，估算方法也多种多样。比较产品的技术规格或者分析产品说明中相同的字词，就有可能判断出两件产品是否相似。但是，对推荐效果影响最大的数据是两件产品被同时购买或浏览的频次。如果发现顾客们有同时购买某两件商品的行为，这两件产品就会被标记为互补商品。如果在同一次购物过程中，顾客们有同时点击两件相似产品的行为，这两件产品就会被标记为替代商品。当顾客查看某一件产品时，他前期的查询、点击和购买数据就会被合并进行分析，并得到替代商品（“顾客在查看本产品后购买的其他产品是什么”）和互补商品（“购买本产品的顾客还购买了……”）等建议。同时，有了顾客点击某件产品并最终购买该产品（或其替代产品）的人数百分比，就可以把这些用户数据转化成有用的决策行为简明程序。

于是，亚马逊在汇总点击和购买数据的基础上开发了自己的商品推荐系统。此外，公司还建立了一个第三方在线销售平台，不仅将仓储空间提供给第三方公司使用，还增加了可供分析的数据来源。亚马逊没有将市场分成几十个细分市场（比如邮购商品目录中常见的“郊区中产阶级妈妈”“快速自提型顾客”等），但是，亚马逊可以做到“将单个顾客一分为十”，及时掌握每个人多变的需求与兴趣。

就其本身而言，数据存储算不上革命性举措。令亚马逊脱颖而出的是，它始终致力于数据挖掘，根据顾客的兴趣、偏好与当前状况向他们推荐商品。但是，过度的个性化推荐有可能吓跑顾客。《习惯的力量》注一书的作者查尔斯·杜希格举过一个十分精彩的例子。一位年轻女子的购物行为触发了塔吉特超市的算法程序，并成了该商场母婴产品定向

广告的发送目标。她的父亲为此勃然大怒，但是几天之后，该女子告诉她父亲她真的怀孕了。塔吉特超市的算法做出了正确的预测。

亚马逊利用顾客与网站互动时产生的所有数据，改变了营销活动。它还通过商品评论，赋予了顾客创造数据的权利。这个实验着重强调品牌控制和产品传播两个方面，使传统营销发生了颠覆性变化。顾客急切地希望同其他顾客分享体验，相较于制造商、推销人员和卖家的介绍，他们通常更相信其他顾客的评价。如果有多位顾客给某个产品的评级较低，即使这件产品深受专家或者公司员工的青睐也无济于事。允许顾客发表评论的做法还有助于让亚马逊的所有产品获得更多的曝光机会，顾客也有机会了解更多顾客的看法，而不会只受某一个人的影响。最终，亚马逊裁撤了网站编辑人员，安排人手开发算法，并将最有用的顾客评论展示在产品页面的顶端。加大对技术与数据的投资以改进顾客的购物体验，这种做法的效果优于在综合管理上增加投入。

亚马逊的数据挖掘改变了10亿人的购物习惯。2015年，美国零售业有近半数的购买活动是从登录亚马逊网站搜索产品开始的（不管这名顾客最终是在哪里购买了该产品）。

开车时，我们无须对内燃机的复杂原理了如指掌。同样，在寻找与我们的兴趣、需求相吻合的产品时，我们也无须了解亚马逊开发的复杂算法的所有原理。我们更需要掌握的内容是机器的基础结构和安全操作规程。当我们利用更多的数据源和检测装置创建、分享数据时，我们可以袖手旁观，任由他人制定使用条款（翻过20多个页面之后，开开心心地点击确认按钮），也可以主动参与建立新的规范。我们可以把社交数据挖掘程序视为神秘的“黑匣子”，也可以钻研数据知识，并要求数据服务商提供几种有效方法，让我们可以对数据挖掘施加影响，以确保在把个人数据交给他们的同时，我们能够获得等价的回报。

你的数据有什么价值？

在日常生活中，很多决策活动（例如，在亚马逊网站上买什么，晚饭去哪里吃，怎么去那里）已经对社交数据产生了依赖。随着生活中创造社交数据的领域越来越多，我们对数据挖掘的依赖性也将不断增加，用它帮助我们做出某些重大的人生决策，包括选择伴侣、工作地点与工

作方式、药物治疗方案、学习方式和学习内容等。

在很多情况下，只有将我们创造的数据与其他人创建的数据放到一起比较，它们的真实含义才会浮现出来。由于可用于数据挖掘的社交数据呈指数增长趋势，以前我们百思不得其解的很多问题，现在也能找到答案了。我们甚至还有可能提出一些有意义的新问题，而之前我们根本想不到这些问题。

算法可以发现人类不借助计算机就无法发现的规律，这些规律有助于我们的决策。与数据服务商共享数据的好处，取决于数据挖掘的产出对于我们的决策活动（包括为我们自己和亲友进行的交易谈判、购买商品和服务、申请贷款、找工作、接受医疗服务、接受教育、改善社区安全状况、改进公共服务等）有多少帮助。

一旦我们开始思考数据公司的产出可以给我们带来哪些好处，就说明我们已经发生了显著变化。我们不再纠结于一些老问题，比如，公司与政府如何、何时、为什么收集我们的"数字化排放"（也就是我们每天创造的数据）。有人认为，已经有太多数据被收集了，因此个人的最佳选择是不要过多分享自己的数据，或者在有人收集你创造、分享的数据时向他们收费。然而，我们对输入的关注导致我们错过了潜在的好处。我认为，我们在贡献自己的原始数据之后，应该要求更有价值的回报，而不应只满足于那一点儿金钱。我们应该要求数据公司给我们留有一席之地，让我们有机会依据容易理解的公平合作条款，对数据挖掘的产出施加影响。

首先，请思考原始数据与精炼数据在价值上有何差异。如果我在谷歌搜索框中输入我的姓名"Andreas　Weigend"（安德雷斯·韦思岸），谷歌就会告诉我，"大约有122　000"个页面包含单词"Andreas"（安德雷斯）和"Weigend"（韦思岸）。任何人都无法用手动的方式筛选出这么多的网页。如果以非常快的速度点击这些网页并查看，每个页面需要5秒钟，那么看完这些页面需要整整一周的时间。这显然是行不通的。因此，我们只能依赖谷歌为我们提供的按序排列的搜索结果。谷歌有可能会将最近提及这两个单词的网页放在第一位。如果我感兴趣的是关于我的最新消息，那么这种排序就非常理想；但是，如果我要找的是我几年前上课的视频，这种排序就会让我大费周章。另外一种排序方法是计算我的姓名在网页上出现的次数，把出现次数最多的页面设定为相关性最高的搜索结果。如果我希望从一堆文章中找到我的观点被引用最多的那

一篇，那么这种排序法或许有一定帮助。但是，如果我搜索的不是自己的姓名，而是“价格便宜的苹果平板电脑”，搜索结果有350 000个，那么这种排序法是否有用呢？善于骗点击率的人（这样的人为数不少）经常会加载含有热门搜索项的页面，以至于我经常不得不在铺天盖地的搜索结果中，费力地点击一个个链接，希望能找到一个真正有用的页面。

为了提高搜索结果的有效性，谷歌在评估页面适用性时，不仅会搜索关键字词，还会从多个数据源获取数据。首先，谷歌工程师会根据其他页面指向该页面的次数，对相关页面进行排序。这种导入链接对人们的意图有一定的预示作用。在人们发现导入链接可以对页面在搜索结果中的排序产生重要影响之后，“搜索引擎优化”这个领域（包括声名不佳的链接工厂）便应运而生。谷歌的算法必须适应这个变化，学会辨别哪些导入链接是真正感兴趣的个人创建的，而哪些是为了谋利而创建的。现在，除了网络的链接结构之外，谷歌还对人们通过搜索访问的页面以及在这些页面上的驻留时间（在退回搜索结果列表并点击下一个链接之前，停留在该页面上的时长）进行了跟踪，并收集了近20年的相关数据。如果有多名访问者点击某个页面，但在粗略浏览（即所谓的“短暂点击”）之后就迅速离开，谷歌就会将这个页面从相关排序中剔除。但是，在谷歌搜索结果中排序靠前，只能说明人们对该页面的关注度较高，并不能保证页面上的信息一定符合用户的要求。

想一想，谷歌每天执行多少次搜索命令？脸谱网上每天发出多少张照片？培养数据素养，你必须掌握的一个基本技能是了解哪些数据合乎情理，哪些不合情理，哪些是假的。判断数字是否准确并不重要，具备数据素养意味着你可以看出哪些数据一般来说是有道理的，哪些数据的数量级可能出错了。物理学家在评估数据时，往往会用数量级或十分之几这样的术语来表示。他们会说，使用谷歌和脸谱网的人数达到10亿这个数量级，因为他们确定具体用户人数超过1亿，但不到100亿。接着，他们会假设普通用户每人每天进行的搜索次数是10这个数量级，因为这个次数肯定在1与100之间。每天在脸谱网上发出的照片数是1这个数量级，因为具体数量在1与10之间。这样一来，对于这两个常见的社交活动数据，我们可以给出一个用数量级表示的估算结果：谷歌搜索引擎每天执行的命令数是100亿这个数量级，脸谱网每天发出的照片数量是10亿这个数量级。

如果你承认社交数据点每天正在以几十亿的数量不断增加，你就会

明白，从金钱这个角度看，你的原始数据流价值不高是合情合理的。它与个人数据可能带给你的情感价值无法比拟。你把爱犬的照片发在脸谱网上，对它感兴趣的人可能不超过100个，也就是说，还不到脸谱网用户总数的0.000　01%。只有汇集数百万人的数据并加以分析，才可能发现有价值的相关性或规律性。剔除某一个人的数据，数据公司用剩下的数据仍然可以得出相同的结论。数据被剔除的个人将一无所获，而数据公司则几乎没有任何损失，只不过在10亿人的数据总量中减去了一个人的数据。

此外，输入的数据与上传到脸谱网上的照片不同，不一定都是离散数据。单个数据点可能像掉进海底的一颗石子，甚至是一粒沙，尽管难以发现，但却独立存在。它也可能像一滴墨水，在显微镜下观察，这滴墨水会在水中不断扩散，最终与水浑然一体。培养数据素养，你还需要知道何时你的数据独立存在，何时与总体数据融为一体。我在前面说过，在亚马逊网站上，点击一件产品的行为与点击另一件产品或者购买某件产品的行为是有关联的。如果顾客不希望这次购买行为出现在他的购物历史记录中，那么他可以删除这一条记录。但是，在亚马逊的产品推荐系统中，由于点击行为与顾客本人并不是关联的，因此他不可能将这次点击行为从该系统中删除。这个特点再次说明石油提炼与数据挖掘之间具有相似性。在某个时间点之后，单个油井产出的石油就再也不能从炼油程序中分离出来了。

以这种方式从质和量两个方面理解数据的意义，是我不赞成个人数据收费的部分原因（虽然不是全部原因）。自2013年出版《互联网冲击》[注]这本书之后，微软研究院的学者杰伦·拉尼尔（Jaron Lanier）就积极地扮演起啦啦队队长的角色，为通过“小额支付”这种形式实现数据补偿的做法摇旗呐喊。他最欣赏的一个案例是谷歌提供的文本翻译服务。他请大家思考一个问题：谷歌凭什么将所有广告收益悉数收入囊中，而那些针对翻译服务提出建议、修正错误，从而改进谷歌算法的人却一无所获呢？每一条建议、每一次修改，都会让谷歌的文本翻译模型有所改进。即使新的意见是直接复制前人意见的结果，也同样能做出贡献，因为意见被复制之后，模型就会更加重视这条意见。

拉尼尔笔下的那些做出贡献的人，其实已经得到了回报。这些人很有可能也使用了谷歌的文本翻译服务，因此他们已经得到了回报。只不过他们得到的不是金钱，而是更完善的数据产品和服务。

接下来，我们考虑脸谱网上每天产生的一部分数据。如果你上传爱犬的照片，你就创建了数据。但是，如果你上传的是生日派对上一群好友的合影呢？你拍摄了这张照片，然后上传，但是对脸谱网来说，这张照片的商业价值取决于它带来的流量，以及它携带的精炼数据，因为这些数据可以反映通常蕴含在人际交往之中的各种关系和利益。你分享这些数据，应该得到由其带来的全部回报呢，还是与照片里的其他所有人均分回报呢？如果有人添加了评论或者标签，这张照片就变成了供朋友们观看的生日活动的一部分，此时，你又该分到多少回报呢？这些数据的量非常大，对数据公司的价值也非常大，可以改进它们的服务，确保它们有利可图。拉尼尔没有讨论这些数据，也许是因为他认为它们并不是值得人们掏腰包的有“独创性”的内容吧。但是，数据公司每天都以这些数字痕迹为主要原材料，生产出我们无法摆脱的各种数字产品和服务。

如果你所有的点击与搜索、你加的所有关注与标签，都要求数据公司付费，还要考虑其他所有人接触这些数据、添加新数据等因素，那么可以肯定的是，当用户使用搜索结果、推荐意见和结果排序等服务时，数据公司也会要求用户付费。开发算法需要投入，而数据分析需要开发特殊的算法，才能为所有数据赋予属性和价值，包括数据价值随时间变化而变化这一属性。

拉尼尔倡导的小额支付方案之所以流产，原因不只在于解决属性问题所需要付出的成本和面对的困难。首先，我们考虑一个非常简单的原因。如果脸谱网不给股东发红利，而直接将所有利润（2015年的利润约为35亿美元）全部分给用户，那么每名用户当年大约可以分得3.50美元。请问，一个没有任何限制的交流平台在你心中的价值是不是高于一杯卡布奇诺咖啡？如果你回答是，那么已经有人为你的数据“付费”了。其次，在很多情况下，为了得到数据产品和服务，你必须先提供数据，比如，使用优步拼车应用时就需要提交你的位置数据。你可以下决心不再向数据服务商提供免费数据，那么你也得不到它们提供的免费产品和服务。再次，有很多产出（包括产品推荐、了解出租车何时会供不应求等），数据服务商仅需用个人的原始数据即可完成。尽管你的具体数据可能不会改变你看到的结果，但数据服务商却必须要求使用它们产品和服务的所有人都贡献数据。

基于上述原因，我认为你不应因为贡献了原始数据而向数据服务商

索要报酬，而应该要求得到更多的影响力，以便在如何、何时和为什么分享数据，你的数据的用途，你能得到什么回报等问题上施加影响。最成功的数据服务商会向你明示，你贡献的数据有利于它们为你提供质量更好的数据产品。我们整个社会都在喋喋不休地讨论，是否应该采取措施限制各机构利用我们的原始数据的问题，但却很少考虑应该要求数据服务商提供哪些工具，以便增加透明性和主动性这个问题。

数据服务商不会将我们变成一串用来买卖的数字，至少，它们不必如此。如果本书能对你有所启发，我希望你能明白这样一个道理：社交数据可以为你的决策提供帮助，而不只是帮助某个特大企业发起一场目标更明确的广告运动。我相信，你创建的数据必然符合你的特点，你做出的决策也必然符合你的特点。这就是你的数据的价值。

老虎机与挑剔的相亲者

数据挖掘程序还需要在探索与开发之间取得平衡。看到这句话，你的脑海里也许会浮现出幽暗、肮脏的街角。其实，我希望带你参观灯红酒绿、吃角子老虎机随处可见的拉斯韦加斯长街。在人工智能领域（也就是让计算机软件通过输入数据进行学习的那个领域），“老虎机问题”（one-armed bandit problem）具有非常重要的地位。它描述了一个两难困境：到底应该探索新的选择，还是接受既有的最好选择呢？打个比方，你走进赌场，听说有人在某台老虎机上赢了一大笔钱。你会怎么做呢？通过观察，你发现那台老虎机吐出的钱比其他机器都多，你会不会在接下来的时间里都守在这台老虎机的旁边呢？还是你也会尝试其他老虎机，通过数据找到赢取累计奖金概率最高的那台机器呢？当然，收集数据需要一定的时间。而且，赌场总是要赚钱的，所以根据这个赌博游戏的设定，赌徒总体来说是要输钱的。计算机理论专家说，最理想的情况是你花一些时间观察这些老虎机，然后从中发现规律。但是，你到底应该在赌场每台喧闹的老虎机上花多少时间呢？尽管统计学家有可能给出建议，但你仍然需要做出取舍：是探索新的方案，还是利用既有的最佳方案呢？老虎机问题看似与数据挖掘的产出没有关系，但它在考虑提交给用户的推荐意见如何排序、用户如何选择最适合的推荐意见，以及如何在探索与利用之间取得平衡这些问题上具有非常重要的意义。此时把数据类比成石油，可能会帮助我们更好地理解这个问题。石油地质

学家和工程师在开发油田时，经常需要经过权衡制订出一个折中方案：是投入大量资源和大笔资金，不断增加开采深度，直至油田枯竭，还是调动人力、物力，寻找出油率更高的新油田呢？数据服务商在分配资源时也需要做出决策，以实现投入产出比和效率的最大化。涉及数据时，需要加以管理的最重要资源就是用户的时间。

当谷歌等搜索引擎响应用户的查询需求时，它们不仅会大批量地显示可能与用户的搜索命令相吻合或高度相似的结果，还会列出一些可选方案，即有可能与搜索项有一定相关性的页面。有时候，你想要搜索的是关于某个事物的信息，例如，你在查询框中输入的是“Panthera onca”（美洲豹）这个词。但是，如果你输入的搜索项是“jaguar”（捷豹），搜索结果就不全是与美洲豹这种猫科动物有关的网页，还有与“捷豹”汽车、苹果曾经使用的那套操作系统相关的网页。搜索引擎的算法会依据页面上的字词、页面间的链接，以及人们在页面间的跳转情况，为“jaguar”这个词确定几个含义，并为每个含义选择一批相关的搜索结果，供用户查询。搜索引擎希望这种做法可以确保你找到你想要搜索的目标。

老虎机问题的一个衍生难题叫作最优停止理论，或者“挑剔的相亲者”问题。马丁·加德纳（Martin Gardner）在《科学美国人》杂志的“数学游戏”专栏里第一次提到了这类问题。加德纳版的最优停止理论需要使用写有数字的纸条，数字大小不限，“从小于1的分数到像‘古戈尔’（1后面有100个0）这样的大数都可以”。将纸条的顺序打乱，然后一张张翻过来，直到你认为翻开的纸条上的数字是所有纸条中最大的。随着时间的推移，这个思想实验中的纸条变成了相亲的年轻男子。他每次与一名姑娘约会，都必须做出判断：是继续与其他姑娘约会，还是停止约会，认定（在所有约过会的姑娘中）目前这一位是他最理想的伴侣。这就是现实世界中必须在探索和开发中做出抉择的高风险难题。

出于一些显而易见的原因，交友应用程序或交友网站的用户经常需要面对“挑剔的相亲者”这个难题。早期的交友网站要求用户表明他们在体重、身高或者地理区域等方面的偏好，然后依照这些偏好排列约会对象的先后次序。一名用户点击了一位候选约会对象（我们假设她叫萨姆）的照片，而网站并不知道是什么因素促使这位用户点击了萨姆的照片。是因为萨姆在候选人名单上排在第一位吗？是因为她有黑头发或者戴眼镜吗？是因为照片的背景是大海，而且他对住在海边或者喜欢去海

边度假的人感兴趣吗？个中原因无法确定，但他仍然需要做出决定，是给萨姆发一条信息还是再看看其他候选人的照片呢？传统意义上的媒人通常会努力帮助每一名委托人觅得良缘，但交友网站选择让用户自行决定：他可以要求网站提供更多的候选对象，相似的或截然不同的候选人。

对于如何在探索与开发之间取得平衡的问题，大部分数据服务商都会先观察用户探索新的推荐意见的行为可以坚持多久，他们是否会回头、何时回头，然后依据观察结果做出决定。不过，最佳选择常常取决于用户的近期状况。挑剔的相亲者希望寻觅的有可能只是一时的爱情，所以服务商很难猜出最适合他的约会对象到底是哪一个。系统的透明性原则要求用户能够了解服务商的设定，主动性原则要求用户在一定程度上能够影响这些设定。

MoodLogic公司是我和朋友合伙创立的一家音乐推荐公司。我们为每一名用户赋予了一定的权利，让他们自行确定探索与开发之间的平衡点。具体来说，就是让他们自行选择，是由我们为他们提供与其他们常听的音乐风格比较接近的歌曲，还是推荐全新风格的音乐作品。我们分析用户已有的数字音乐库，然后建立一个模型，帮助他们寻找与其音乐库中歌曲相匹配的歌曲、艺术家、作曲家、乐器组合、节拍和流派等。模型可以预测他们对某首新歌的喜爱程度，以及我们对推荐模型的自信程度。接着，我们让用户在两个设定中做出选择。如果他们选择“安全”设定，就会听到风格比较相似的歌曲；如果他们选择“探索”设定，就会听到风格迥异的音乐作品——根据我们的推测，这些歌他们可能喜欢，也可能不喜欢。把选择权交给用户，用户因此创建的数据又可以为我们所用，帮助我们改进算法。

尽管数据可能取之不尽，但时间不可能用之不竭。人们必须下定决心，做出取舍。社交数据的神奇之处就在于，数据挖掘程序的产出反过来又可以变成输入。

通过机器学习发现错误

人们往往对自己做出的决策充满自信。比较可靠的做法是把可选方案的利弊列成清单（我应该接受另一座城市的工作邀请，还是现在所在

公司发出的与前者相差无几的工作邀请呢？），然后经过权衡，选出与当前状况、目标和风险承受度等更吻合的方案。过去，人们生活在“小数据”社会里，只能通过与亲朋好友、同事或导师交谈的方式收集信息，做出决策。

现在，我们可以登录Glassdoor网站，查询工作满意度排名。Glassdoor网站是一个可以匿名评论工作环境和薪金的平台，有超过40万家公司在该网站上被评论，网站每年新增评论数为50万条。例如，关于亚马逊的职位和求职面试的评论各有8 000条，还有14 000条薪金评论，涉及1 400个职位。与以前相比，人们在选择工作时有更多的数据可以参考，但他们没有足够的时间去阅读、分析8 000条评论，再与现在的工作做比较。哪些评论意见比较中肯，哪些与你正在考虑的这份工作有最强的相关性？某些评论者是否有可能因为一不小心看错了问题而打了一个较低的分数呢？

所有的数据都有可能出错。在小数据时代，负责收集数据的人都兢兢业业，他们会认真核查每一个数据点，用手工的方式剔除、修改大多数的错误。他们检查所有数据的这种行为真的值得称赞，因为有很多关系到整个地区、整个州的决策所依赖的数据，都来自这些规模很小的样本统计结果。如果某个州某个星期申请失业救济的总人数出现了错误（例如，将254误写成2 541），这个错误经过累计之后就有可能影响到年失业人数，进而对政府的经济政策产生影响。美国劳工部劳工统计局为工人建立了一个1万人的调查样本，样本人数与亚马逊员工在Glassdoor网站上发表的评论数属于同一个数量级。

我们可以做出一个合理假设：数据中的错误率与所收集数据的总量之间没有关系。如果我们现在可以用的数据量增加100倍，那么不正确的数据点也应该增加约100倍。但是，核查所有数据并将错误一一剔除，已经是不可能做到的事情了。

不过，数据量的指数增长也为数据错误的指数增长提供了一个解决之道。由于人们在使用数据服务商的产品时会不停地创建新数据，因此算法可以学会判断哪些数据可能是输入错误。如果你输入的是“Andreas Weigand”（安德雷斯·威根德），谷歌就会询问你想要搜索的是不是“Andreas Weigend”，因为其他用户在看到前几个搜索结果之后，有一部分人将查询项改成了“Andreas Weigend”。

数据服务商只需合并不同来源的数据，就可以发现这样的输入错误。2012年7月的某一天，我的手机上出现了一个叫作“Google Now”的应用。这个应用可以扫描谷歌邮箱，查询我的电子机票信息，更新我的航班情况，而且更新速度经常比航空公司还快。真是太方便了！但是，Google Now应用还有更多的惊喜在等着我。一天早晨，我正在收拾行李，准备离开弗莱堡。突然，这个应用通知我必须立即赶往机场。根据时间表，我的航班几个小时之后才会起飞。通常，航班起飞时间的提前幅度不会超过几分钟，因此这个提醒有点儿莫名其妙。不过，我对Google Now应用的信任程度超过日程表，所以我决定立即出发去机场。也许是该应用发现前往机场的路上发生特大交通事故了吧。到了机场，我发现自己输入到谷歌日历上的航班时间是错误的。这款应用忽略了人工输入的错误数据，给我发了一个提示信息。（三年后，我将电子机票放进了我的谷歌邮箱。航班信息被自动添加到我的谷歌日历上。）

这种服务十分有用，我们已逐渐习惯让数据服务商找出并纠正这些错误了。问题是，随着我们创建、分享的数据不断增加，我们是否愿意让它们在我们生活的其他方面发挥类似的作用呢？

数据服务商也必须正确解读数据，辨识信号与噪声。“信号”与“噪声”是统计学术语，前者指相关的数据，后者指随机出现而没有相关性的数据。社交数据非常复杂，信号与噪声的区分因用户而异，因情境而异。你的脸谱网好友上传了一张照片并标签你，但这张照片中并没有你，这是信号还是噪声？答案不确定。如果他本来准备点击安德鲁的姓名，结果不小心点击了你的姓名，因为在他的好友列表中你的姓名就在安德鲁的上面，这个标签就是噪声，是社交数据中的静电噪声，会干扰“收音机”的收听效果。如果他标签你，是为了让你看这张照片，那么，尽管这种行为可能让你讨厌，但它仍然是一个信号——用统计学术语来说，它不是噪声。

通过用户反馈进行机器学习，对于数据服务商改进算法具有重要意义。我说的用户反馈指的不是邀请客户坐到桌旁，填写客户满意度调查表，或者成为调研对象。不断深化与用户之间的“交谈”，可以帮助数据服务商提高数据产品和服务的质量，并增强针对性。你做出的每个选择，都有助于数据服务商调整选项排序。但是，作为用户，为了使搜索结果尽量符合你的期望，你也会不断修改搜索项，不仅会尽力避免输入错误，还会强调你对某个主题不同范畴或不同方面的兴趣。

不过，当你同网站或应用程序交互时，数据服务商提供的选项也会产生限制作用。我认为，如果允许用户对数据挖掘的置信区间产生影响，查询优化的效果就会更明显。我和同事创立的MoodLogic网站在为用户推荐音乐时就是这样做的。由于公司员工在Glassdoor网站上发表的评论大幅增加，网站必须加大数据挖掘的力度，用户才能使用这些数据。模型也许可以判断哪些评论者的看法与某位用户的相关程度最高，其判断依据不只是评论者的职位或者工作地点，还有其他数据，包括职业目标、偏好的工作环境等。不过，无论输入的数据有多少，这种判断都存在不确定性。

具有数据素养，意味着你知道推荐意见不一定都是正确的。做任何决定时都需要在风险与回报之间做出取舍，即使大数据可以将不确定性降低，也需要你做出选择。数据服务商不应该代替你做决定，而应该帮你降低犯错的可能性，让你有能力利用更多的数据。

在数据服务商的帮助下，我们可以使用、分析丰富的历史数据，从中发现规律，预测趋势，尽管难免会犯错。这种认识数据和我们自己的方式，与我们大多数人熟悉的方法有明显的不同。

用数据模型辅助决策

> “给我数据！数据！数据！”（福尔摩斯）不耐烦地喊道，“巧妇也难为无米之炊！”
>
> ——阿瑟·柯南·道尔（Arthur Conan Doyle）

20世纪90年代初，我在施乐帕克研究中心（帕洛阿尔托研究中心）担任博士后研究员，利用超级计算机分析道路交通的规律，以实现行程时间可预测等目标。作为物理学家，我们把交通看成流体，试图找出导致“层流变为紊流”（流动状况由平稳转变为断续不定）的因素。按今天的标准看，我们掌握的数据并不多，因此在创建模拟交通模型时，我们必须做出大量假设。

今天，预测何时到达目的地毫无难度，因为几乎每辆车里都有人携带手机，手机可以实时规划路线。微软在这个领域有一个名叫英瑞克斯（Inrix）的子公司。该公司每天分析一亿多部手机的地理位置数据，了

解它们的目的地（更重要的是，了解它们不会去哪些地方），以推断人与数字产品的运动趋势。英瑞克斯从通信运营商那里获取数据，了解这一亿部手机正在连接哪些蜂窝基站。英瑞克斯完成数据挖掘后，就会将其出售给那些为驾驶员提供导航和路线规划服务的公司，包括佳明（Garmin）、MapQuest网站、福特、宝马等。英瑞克斯还在城市规划方面为政府出谋划策，为修建新桥、加装红绿灯、建设新的公立医院等公共设施选择合适的地点。

英瑞克斯的交通信息更新服务表明，在为决策提供帮助时，数据服务商从众多设备收集并汇聚数据，可以获得1＋1远大于2的效果。基于社交数据的预知系统（anticipatory system），将在包括人际关系、经济、就业、医疗在内的多个方面，对大量决策活动提出建议（甚至有可能给出警告）。这项服务还强调数据解读在数据挖掘过程中具有至关重要的意义。数据服务商建立的数据模型有三种“风味”：描述性模型、预测性模型和指示性模型。描述性模型描述历史特点；预测性模型根据过去的情况预测未来，同时假设与系统的交互以及对系统的操控不会影响结果；指示性模型在分析历史数据的基础上告诉我们该怎么做，才能实现期望的结果。

描述性模型会对数据加以总结，例如根据相似性为数据点归类。在你考虑具体情况时，这类模型可以帮你设置一个标准，使决策有据可依。如果你想了解曼哈顿当前交通发生拥堵的地段，那么你可以跟踪手机的地理位置、查看汽车的移动速度，从而发现拥堵路段。但是，即便是这种比较简单的应用，也需要解读数据。你也许会发现，大量数据表明大都会保险公司大楼附近有静止不动的汽车。但是，这是不是因为大都会保险公司靠近中央火车站，附近还有几个繁忙的出租车停靠点呢？会不会有几名出租车司机在等客，甚至还有几名乘客和他们在一起，因此有为数不少的手机都给出了交通“失速”的错误数据呢？如果你想了解你的商店在某个销售旺季的业绩，算一算销售额就知道了。但是，你还需要将计算结果与其他数据做比较。如果你比较的是去年同期的销售额，那么本地经济的变化情况并没有被考虑进去。因此，正确的做法是比较本地区类似商店的销售额。

我在亚马逊工作期间，和同事们研究过顾客从点击查看商品到最后下单购买的时长。有的数据点明显出错了，因为时长差竟然是负值，购买绝不可能发生在点击查看商品这个行为之前。我们不知道为什么会出

现这样的错误，但还是剔除了这些数据。剩下来的那些数据表明，为数不少的顾客从点击查看到下单购买的时长为8个小时。这个结果太奇怪了！后来，我们发现原因在于亚马逊的计算机时间设置不统一，有的是美国太平洋标准时间，有的是格林尼治标准时间。我们意识到，不同时区间的8个小时的时差，被计入了购买时长。同往常一样，一开始令人激动不已的新“发现”原来只是一个错误。

数据解读是一个迭代过程。举一个例子：一家航空公司希望对潜在的商务舱乘客定向投放手机广告，因此邀请一组数据科学家帮他们寻找定期进出纽约肯尼迪机场的智能手机用户。问题是，进出机场频率最高的不是商务人士，而是机场与航空公司的员工。数据科学家通过观察手机的运动规律发现了这个问题。经常进出机场的一类人是机场工作人员，包括柜台人员、机械师和行李员等。这些人每天进出机场都会按照明确的轮班安排表。家住纽约市的乘务人员进出机场的规律很难描述，但是通过他们利用机场无线网络登录的网站和应用程序可以发现他们的行为特点：他们几乎不会搜索宾馆信息，也不会登录优步应用程序叫出租车，在出机场的路上可能会登录某个交友应用程序。

我们可以把数据挖掘看作预测分析，也就是收集数据并推测未来的情况，包括可能发生的行为与事情。例如，某些城市规划人员就曾利用英瑞克斯公司在一分钟的时间里收集的交通数据，评估某个事件（包括公路事故、工程建设、大型音乐会等）的影响，以便更有效地制订应急方案。一些对冲基金曾利用英瑞克斯收集的大型购物中心和商场周边的交通流量数据，早在这些零售商发布季度销售额之前，就已经做好了买入或抛售股票的决定了。2012年的那个“黑色星期五”，通过分析收集到的地理位置数据，它们成功地预测到当年圣诞节期间的销售额将大幅增加。

与之相似，亚马逊也依赖预测性模型做出商业决策。例如，在节假日等销售旺季，需要增加多少仓库员工与送货员才能完成那些订单。这是一个典型的决策问题：在货品不能被及时送达客户手中与送货能力冗余造成的损失之间如何权衡？亚马逊对送货能力进行了逐日逐城的细化分析。2013年，同几个重要的零售商、货运商一样，亚马逊的预测也出现了错误。很多包裹在圣诞节之后才送达，令顾客大为光火。亚马逊对导致错误的根本原因进行了分析，然后修改了模型，以便更准确地做出预测，有针对性地分配资源。结果，到2014年，公司不仅可以提供商品

包邮服务，保证商品在12月24日前送达，而且最后购买期限较前一年还延后了两天。

由于很多数据服务商还从事商品推荐的业务，因此你必须小心提防排序标准与你的利益不一致的情况。1960年，最早的大型数据开发计划之一，Sabre（半自动业务研究环境）全球订票系统，作为美国航空公司的一个项目正式启动。为开发这套系统，美国航空公司投入大量资源。1976年，Sabre系统成为旅行社的办公软件，其他航空公司的航班信息也被添加进去。美航研究机票预订数据后发现，旅行社最喜欢选择显示屏上排位靠前的航班，如果航班列表超过一页，那么出现在第二页及之后的所有航班它们几乎都不会考虑。于是，美航对算法进行了修改，让排序结果对自己的航班更有利。旅客不知道订票系统提供给他们的前几个选择方案有不公正的问题；而旅行社是拿佣金的，因此它们帮助顾客寻找廉价航班的动机并不强烈。不过，与美航构成竞争关系的纽约航空公司与美国大陆航空公司发现，它们的航班信息被淹没在搜寻结果中，即使它们开设新线路、推出打折机票，效果也不明显，而这两个方法本来可以提升航班的排位。随后，美国国会对此展开了调查。1984年，美国政府下令停止这种不公正的做法。

如果数据服务商的用户不是中间商，而是终端用户，那么他们更有可能关注这些建议与用户偏好之间的契合度。这样一来，这种操纵排序结果的行为就更难禁止了。我曾帮助Agoda（安可达）开发了一套推荐系统，Agoda是一家位于曼谷的宾馆预订网站。开始时，我们以为公司评价宾馆好坏的最佳标准应该是公司从中获取的利润。如果某家宾馆愿意支付优厚的佣金，为什么不把它放到推荐名单的前列呢？或者说应该依据旅客的喜好来排序？有的顾客看到Agoda依据获利多少排列出的结果，可能会直接预订房间，但后来他有可能会后悔。还有的顾客看到排列在前面的几家宾馆之后，认为Agoda无法提供他们喜欢的宾馆，便决定通过Agoda的对手网站来预订房间。从长远来看，把Agoda的利益与旅客的利益捆绑在一起，才是一个更好的选择。

数据处理的终极阶段是规范性分析——从你那里收集数据，然后帮助你决定如何改变现状，以实现理想的结果。美国国家航空航天局（NASA）的登月计划就为数据分析提供了一个经典范例。为了将尼尔·阿姆斯特朗和美国国旗送上月球，NASA的工程师们必须分析连续数据流，以了解登月舱的空间位置。空间工程师需要完成的工作不只是概述

这些数据（描述性），也不只是预测登月舱与月球表面发生碰撞的时间与地点（预测性），还需要在接收到代表登月舱状态变化的数据之后，确定接下来应该采取的行动，以增加将人成功地送到月球上的可能性。每台登月舱喷气推进器点火之后，他们都会了解推进器对登月舱运行轨迹的影响。接着，他们需要预测这台推进器何时需要再次点火、如何点火，以及点火后运行多长时间，才能帮助他们实现目标。

培养数据素养，你必须知道描述离不开假设，不确定性是预测的固有属性，反馈信息对于规范性分析具有极其重要的意义。数据服务商根据谷歌搜索历史记录将你归入某个市场类别，这种做法是否合理？数据服务商判断求职者是否适合某个职位的唯一依据就是他在领英网站上的联系人，这种做法是否可以接受？如果数据服务商只根据某人发在脸谱网上的饭店就餐信息，就建议她调整运动和养生安排，这种做法是否正确？

实验！实验！实验！

数据服务商的工作不只是描述、预测和进行规范性分析，还需要做实验。事实上，只要你上网消费，无论是登录亚马逊购买现代优秀的文学作品、上Zappos网站买软皮平底鞋，还是去Match.com婚恋网站寻找伴侣，你都可能成为某个实验的实验对象。数据服务商需要通过A/B测试来做实验，以便为用户提供更优质的产品和服务。

科学上的因果关系是通过实验确定的。改变实验组的一个自变量，而对照组的这个自变量保持不变，然后比较两组实验对象的反应。A/B测试通常始于某个问题，例如，“如果想销售出去更多的雨伞，那么我应该储备红色的伞还是蓝色的伞？”这个问题看似简单，但在设计A/B测试时会遇到很多难题。雨伞销售商可能会在商场的某个角落设立柜台，在试营业的第一天仅出售红色的伞，而在第二天只出售蓝色的伞，以确定选择哪种雨伞。他甚至可能会把实验时间定在两个连续的周一，因为这个地区的上班族在周一出门时比较仓促，因此更有可能忘记带雨伞。但是，尽管他对地点和时间这两个变量进行了控制，却没有考虑另一个变量：是否会下雨。第三个变量在人们决定是否购买雨伞（包括红色的伞和蓝色的伞）方面具有最大的影响力。

数据服务商需要考虑的变量远多于雨伞销售商。亚马逊在设计页面时，每做一个决定都会使用A/B测试。例如，首页上的搜索框尺寸多大比较合适？结账对话框应该出现在显示屏的左侧还是右侧？无须二次点击即可看到的产品说明多长为宜？谷歌曾通过A/B测试选择蓝色广告链接的颜色深度，一度成为美谈。据谷歌公司内部人士透露，从50个选择中脱颖而出的最终方案，每年为公司多带来2亿美元的广告收益。

描述性分析为人们发现“自然实验”创造了条件。所谓自然实验，是指某个条件因为偶然因素或错误（也就是说，不是实验设计的安排）而发生变化的情况。自然实验的效果是可以观察到的，例如，软件在首次展示中暴露出一个缺陷。法国亚马逊的开发人员出于某个原因，忘记在结账时计入货运成本。这个错误发生后，在短时间内订单数出现“井喷”现象，这也给了亚马逊一个灵感：商品包邮可以增加销量。

预测是科学研究的一个核心内容。科研人员建立一个有预测功能的模型，然后开展实验，检查实验结果是否与预测相吻合。如果不吻合，科研人员就会修改模型，再一次做实验。

社交数据领域中最令我感兴趣的是涉及规范性分析的实验，这类实验允许用户修改某个参数，然后观察产出或结果是否会发生变化。利用数据及时发现交通拥堵路段的数据服务商可以向驾驶员发出预警，提醒他们道路堵塞导致预计行程时间有所延长，或者建议他选择另外一条比较顺畅的路线。然而，如果大多数驾驶员都选择相同的替代路线，就可能导致新的交通拥堵。数据服务商可以提供多条路线，并告知驾驶员该地区已经有多少名驾驶员选择了某条路线，帮助驾驶员决定他们是否应该选择另一条路线。数据服务商也可以利用这些数据，预测在接下来一段时间里是否有可能发生交通拥堵，并通过调整红绿灯的时间预防拥堵发生，优化路况。

我的前同事罗恩·柯哈维（Ron Kohavi）是一名优秀的A/B测试专家。2005年，罗恩离开亚马逊，帮助微软组建了规范性分析实验团队。为了制定在线测试质量评估的基本准则，这支团队在大约20个网站上进行了数百次A/B测试。这些测试让罗恩懂得了一个道理：“获取数据并不难，但获取可靠的数据却非常难。”我由衷地赞同这个观点。这句话也可以套用到数据挖掘上：提供推荐意见并不难，但评估推荐意见的好坏却非常难。

在网站上进行A/B测试时，错误几乎防不胜防。首先，某些网站有15%~30%的网页浏览量是由网络爬虫完成的。机器人进行的这些访问必须辨识出来，并从人类进行的访问中剔除出去，除非数据服务商是在为这些机器人完成优化程序。

在将用户分成实验组和对照组时，我们可能会受到诱惑，放弃随机原则，选择某个看似更有效的分组方式。不过，尽管有很多种非随机抽样的方式都可以取得不错的初期效果，但是这种做法会影响实验结果，使数据分析蒙上污点。比如，如果用户定期清理电脑缓存，那么在不同的实验阶段，她可能会被分配到不同的组。在某些实验中，用户点击某个链接，就会被引导至某个实验组版页面或者对照组版页面，因此分组情况可能与她当时所浏览的网站有关。如果气象频道正在不间断报道可能来袭的飓风，而用户正好在浏览气象频道，那么他点击雨伞广告的可能性会不会有所增加呢？如果不是随机分组，结果就会有偏差。

此外，科学家们还会考虑没有被纳入实验但可能影响用户行为的那些变量。如果某个版本的软件仅提供给某一组实验对象使用，那么它的缺陷就有可能导致实验结果出现偏差。软件在不同平台上的运行情况也会造成偏差。在整体人口中，使用苹果手机接入网络的用户与使用安卓手机接入网络的用户并不是两个独立同分布的变量。实验可能会告诉我们，苹果手机用户访问某个网站的频次高于安卓手机用户，但是，造成差异的原因其实是手机软件（而不是用户群），因为苹果手机默认的页面刷新速度可能比安卓手机高。找出可能符合这些条件的变量，然后展开调查，这是数据侦探们日常工作中的一大乐事。

早在网络问世之前，企业请顾客体验产品和包装的测试活动就已经有几十年的历史了。测试活动的创新点在于，可以迅速收集到反馈信息（包括数据服务商提供的信息），并用到产品和服务的开发工作中。过去，创意和效果之间的时间差长达数月。现在，我们生活在一个全连接世界之中，获得反馈信息的时间已经缩短到几分钟了。这与医学实验的时间精度大不相同，在药物的配方发生改变后，也许需要几周、几个月、几年甚至几十年的时间，效果才会显现出来。

随着社交数据越来越多地应用于解决问题和做出决策，数据服务商会把触角伸入医疗、教育等关乎我们生活的重要领域，开发并提供数据产品和服务。我们必须考虑应该进行哪些社交数据实验，以及我们可以信任哪些实验结果。在哪些情况下，我们收集一个小时的数据就足够

了？在哪些情况下，我们需要收集一天的数据？在哪些情况下，我们做实验时最好采用较大的时间单位呢？比如，在教育领域，这些问题就没有明显正确的答案。在利用A/B测试检验教学改进措施的效果时，我们先要明确教学改革的目标，再测试相关数据。前文提到过，提出推荐意见并不难，但评估这些意见的好坏却非常难。

然而，我们不能因此感到害怕。用社交数据生产的某些数据产品和服务在20年前还是不可想象的东西，现在却已经走进了我们的生活，并且像水、电等基础资源一样不可或缺。可以说，我们所有人都已经享受到社交数据实验的胜利果实了。社交数据还有不计其数的创新用法，但受到资金预算、社会规范和独创性的限制。为了享用大型数据服务商的劳动成果，我们必须欣然接受我们都是实验对象这个事实，并且鼓励数据科学家通过实验，为我们的决策活动提供建议。我们不应该继续在黑暗中摸索。

基于以上原因，我认为在获得新的数据权利之前，我们必须加强对数据服务商的管理，我们必须更好地理解社交数据的三个来源：我们点击鼠标的行为、我们的人脉，以及我们所处的环境。我们将会发现，原始数据的这三种来源将挑战很多既有的社会规范，包括一些根深蒂固，通常还涉及情感的社会规范。我们如何确定个人身份？隐私权在多大程度上是一个幻象？好友的含义是什么？如何判断我们该信任谁，何时可以信任他，为什么信任他？我们和环境之间的相互影响程度有多深？如果我告诉你，所有这些问题的答案都可以从你在谷歌上的搜索行为、你在脸谱网上的交友行为，以及你手机的传感器中找到，你也许会大吃一惊吧。

---

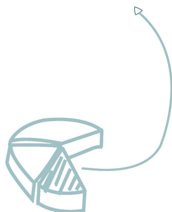
1. 《习惯的力量》简体中文版由中信出版社于2013年3月出版。——编者注
2. 《互联网冲击》简体中文版由中信出版社于2014年5月出版。——编者注



第2章
数字身份与真实身份

数字时代隐私权与诚信的博弈

你创建的数据里有你的影子吗？



行为反映身份，行为决定命运。

——罗伯特·穆齐尔（Robert Musil）

我上学时学的是物理学，现在从事社交数据实验的很多人也是如此。这并不奇怪，因为我们浏览网页、使用手机等行为留下来的数字痕迹与粒子探测器捕捉到的路径及计数非常相似。事实上，从事实验性粒子物理研究的经历为人类完成电子商务实验奠定了坚实的基础。

高能物理领域的研究人员不可能真正看到粒子的真实身影，唯一的办法就是建造探测器，观察粒子之间的相互作用。物理学家通过分析这些相互作用，推测粒子的属性。上大学时，我曾在日内瓦附近的欧洲粒子物理研究所（CERN）研究过一个气泡室实验的数据。粒子一进入气泡室，就会与接近沸点的液体发生反应并形成气泡。这个实验的目的是测量这个微小气泡的运动轨迹与半径，有了这两个数据就可以计算粒子的电量和质量。这种间接测量的原则适用于几乎所有的粒子物理实验。没有人可以看到希格斯玻色子，但是，由于人们观察到了一些间接的痕

迹，因此大多数物理学家都确信这种粒子是存在的。

同粒子一样，只需观察一个人同其他人和物体之间的交互，以及他对这些人和物体的重视程度，就可以知道他的性格特点。这与罗伯特·穆齐尔的小说《没有个性的人》中的男主角——性格随周围环境变化的数学家——有点儿相似。我们留下的数字痕迹中包含大量关于我们自己的信息，对我们的隐私来说是一个挑战。数据服务商还可以根据这些数字痕迹，观察、分析我们的行为举止，然后预测我们的行为与兴趣爱好，虽然我们可能根本不认识他们。

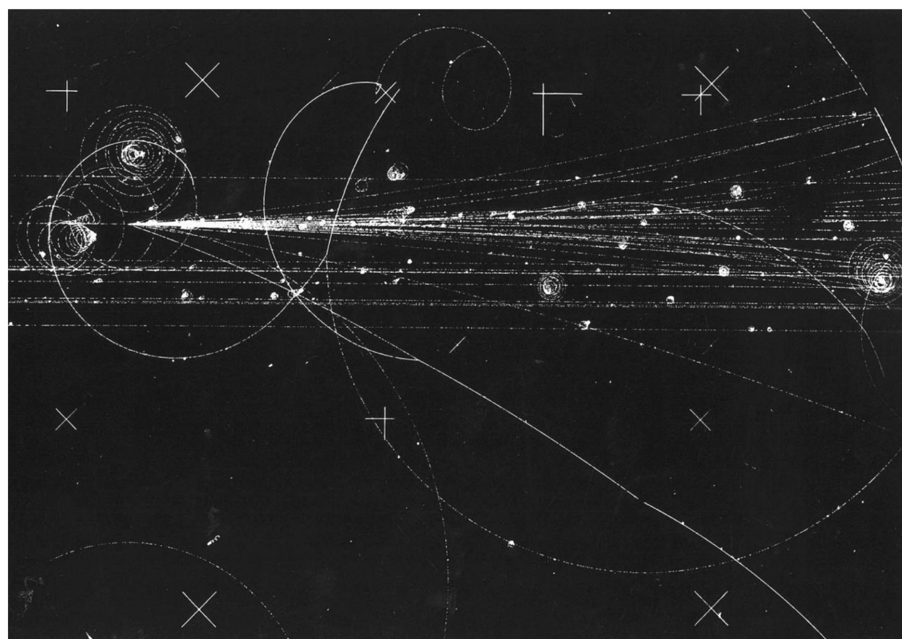


图2–1 在费米国家加速器实验室的气泡室实验中产生的26个带电粒子的质子运动轨迹
资料来源：本次转载获得了费米实验室的许可。

新的社交数据平台为我们展现自己提供了前所未有的良机，因此我们积极地创建并分享我们的个人数据。在“打理”自我时，我们不再局限于变换发型、改变穿着打扮等所谓的炫耀性消费。现在，我们可以轻松地改变自己的数字身份，包括头像照片、假名等。我把它称作“炫耀性交流”。互联网刚刚兴起时，我们一度以为不公开姓名可以赋予我们新的自由。由于性格方面的资料有被人恶意利用的风险，因此我们自然希望建立相关的规则来保护自己的隐私，但是不公开姓名同样会对我们利用这些数据造成障碍。

面对个人数据可能遭到恶意利用的问题，许多人的第一反应是建立

更有力的隐私保护机制，因为他们认为个人隐私权是不可侵犯的。不过，很多隐私保护措施其实增加了规范用户言行的难度。透明性与主动性原则有利于用户从个人数据中获取价值，但是公信力缺失却与这些原则背道而驰。在社交数据泛滥的世界里，旧有规则已经不适用了。无处不在的数据创造与数据传播在带来经济效益的同时，还要求我们搭建新的框架，确立新的理想。

不过，在展望未来之前，我们先仔细研究一下隐私权营造的幻象。隐私权是技术于不久之前催生的一个概念。至少在当下的世界里，这个概念的任务并不是为生活在社交数据时代的人提供保护。不能因为隐私权是对100年前的老问题做出的一个积极回应，我们就义无反顾地投身到保卫隐私权的战争当中。

隐私权简史

纵观历史，人类大部分时间都生活在公共环境之中。我们不仅与家人分享生活空间，关系稍远的亲友也经常与我们一起围炉而坐，对我们的生活起居也不无了解。抵近观察使人们宛若“近邻”，破坏社会规则的人将遭到流言蜚语的无情唾弃，情节严重的话，甚至会遭到流放或者更严重的惩罚。

烟囱很可能是催生隐私权的第一项技术。17世纪，烟囱成了欧洲家庭的一个常见特征，更多的家庭利用围墙和门窗阻挡亲友窥探的目光，把家庭变成隐秘的领地。与此同时，农业上的一连串创新技术改变了人们的谋生方式。至18世纪中叶，食品生产的速度已经超过人口的增长速度，尽管人口呈快速增长趋势。很多人迁居城市，产业革命在城市里掀起了第一波兴建工厂的浪潮。

城市居民过着一种隐姓埋名的生活，他们关门闭户，将陌生人拒之门外。人们通常把壁炉设计得非常深，以便把锅架到火上。但这种设计十分低效，烟会在狭小的空间里越积越多，这是拥有隐秘空间不得不付出的代价。18世纪40年代，情况终于得到了改观。当然，这要归功于本杰明·富兰克林。他设计的“宾夕法尼亚壁炉”采暖效果更好，还能有效地将烟经由烟囱排出。这样一来，人们就可以关上房门，而不用担心窒息了。在人们的心中，家变成了至圣所，披上了隐私与安全的外衣。

尽管家庭生活越来越讲求隐秘性，政治生活却反其道行之。在我们研究民主的早期实验中，投票显然是社会活动。毕竟，这项活动的目的是增加普通公民的言论自由。据哈佛大学历史学教授吉尔·莱波雷（Jill Lepore）介绍，美国建国后的头100年里，人们都是通过举手或者在房间一侧排成一排的形式公开投票的（艾奥瓦州的总统选举党团会议至今仍采取这种形式）。莱波雷说，“无记名投票”被视为“胆小懦弱、不敢见人、卑鄙可耻的行为”，会破坏选举的公开性和直接交流，而众多思想家认为这两项内容恰恰是民主政治的重要组成部分。例如，从19世纪50年代开始，英国哲学家约翰·斯图尔特·穆勒（John Stuart Mill）就认为，无记名投票容易受到“私利”的影响，因此“通行的做法不应该是无记名投票，而是公开投票”。绅士在投票时应该以公众利益为重，而不应只想着一己之利。要实现这个目的，最好的做法就是让他在公共场合投出负责任的一票，即增加投票活动的透明性。

在只有白人才享有选举权的日子里，纸质选票（当时最先进的技术）被视为高人一等的象征。使用纸质选票，要求选民必须识字，但并不是所有人都有文化。不过，纸质选票最终还是击败了挥舞双臂或摆动身体等投票方式，成为更稳定的计票媒介。最初，选民需要把自己的纸质选票带到投票站。后来，为了便于投票，允许选民提前填好选票，或者请人代劳。有人愿意掏钱印刷选票供其他选民使用，事实上，是他们想操控投票结果。当然，还有其他一些公然的以谋取私利为目的的选举策略。例如，有的政党在供忠诚党员使用以及随机发放给路人的纸质选票上，仅列出了本党的候选人。人们选用纸质选票，不是因为它有利于保护隐私，而是因为它的永久性——可以被重新统计的永久性记录。

1856年，第一张政府印制的无记名选票出现在澳大利亚维多利亚市的选举中；英国引入这套投票系统花了一代人的时间；从19世纪80年代末开始，美国有多座城市、多个州进行投票方式的改革。但是，新的投票方式取得的效果非常有限。19世纪中晚期，美国选民前往投票站的比例普遍达到80%，但是自此之后，再也没有达到这个比例，或许是因为放弃投票的社会成本非常低吧。

就在无记名选票越来越受欢迎的时候，波士顿的两名律师正在为争取新型“隐私权”而据理力争。1890年，前法律事务所合伙人塞缪尔·沃伦（Samuel Warren）与路易斯·布兰代斯（Louis Brandeis）在《哈佛法律评论》上发表文章，严词斥责日益严重的侵犯个人隐私权的现象。这

被普遍认为是人们第一次使用“隐私权”这个词。那么，谁是侵犯者呢？答案是“近期的一些发明创造与经营方法”，包括照片和急于刊登流言蜚语以增加发行量的报纸。同许多发明一样，这两位律师提出“隐私权”这个概念的目的也是为了解决个人问题：不久前，社会新闻栏目瞄上了沃伦及其家人，直言不讳地对他们进行了不利的报道。他们生活的那个年代，与脸谱网每天发出10亿张照片的今天，显然是不一样的。

唉，他们真是高明的律师啊！因为迫切希望把妻子和女儿从社交尴尬中解救出来，他们把抑制自我表现的欲望和控制他人、不让他人随意议论他们的权利混为一谈。在健全的民主社会里，不允许强迫任何人（甚至包括犯罪嫌疑人）透露自己的想法和感受。同某人分享秘密时，很可能隔墙有耳。法律无法阻止这种轻率的言行，但是，在“公众”可以得到更大好处时，社会规范也许有办法让人慎言。工程技术领域通常认为交流的目的是传递信息，但是，脸谱网创始人马克·扎克伯格（Mark Zuckerberg）认为，把获取信息视为目的，其实是为人们之间的交流提供一个借口。

100年前，当布兰代斯被批准进入美国最高法院时，他迫切希望推广的就是隐私权，并将隐私权同美国人坚信不疑的个人自由联系在一起。我们以梅耶诉内布拉斯加州案为例。这个案件争论的焦点是内布拉斯加州是否有权判定教师用德语上课是违法行为。当时，第一次世界大战刚刚结束，反德情绪还很强烈。最高法院的大多数法官坚信，无论家住何处，人人都“有权签订个人契约，有权从事任何一般性职业，有权获得有用的知识，有权结婚、建立家庭和抚养孩子，有权遵从自己内心的意愿去信仰上帝。这些权利还是在普通法历史上被长期承认的私人特权，是自由人有秩序地追求幸福之根本”。按照法律，任何侵害隐私权的行为也是侵犯个人自由权的行为。

我们的个人选择或决定似乎非常安全，不会沦为他人窥探和评判的对象。但是，这或许是我们在自欺欺人吧。随着发现信息的手段和网上交流的工具不断增加，我们对隐私权的缺省假设将会发生显著变化。

从密不透风到公之于众

1996年，拉里·佩奇（Larry Page）和谢尔盖·布林（Sergey Brin）

通过网页之间的链接结构，向建立网络搜索引擎这个目标进发。他们所能利用的唯一工具就是公开数据，谷歌搜索的所有网页都是公开的。某个人做好网页后，把它放到互联网上，让其他人阅读，或者通过链接连到这个网页上。

开发谷歌算法、建造服务器网络需要大笔资金，拉里和谢尔盖知道，最好的办法就是在用户搜索的基础上出售广告空间。广告商挑选出它们认为潜在顾客感兴趣的关键词、短语和产品类别，然后掏钱“购买”，而且广告商的投入立刻带来了回报。与平均数据相比，广告商使用谷歌个性化广告方案（包括有产品相关信息的网页广告）之后，网站的点击人数增加了三倍。用户的搜索数据是非常宝贵的商品，通过这些数据，可以观察和分析人们对哪些产品感兴趣。

2004年4月，谷歌推出了谷歌邮箱服务，为了解用户关注点提供了新的数据来源。谷歌通过分析用户电子邮件的内容，来确定邮箱服务界面上的广告内容。到目前为止，大多数人都认为电子邮件与普通信件是一回事儿。普通信件被密封在信封中，目的是只让信封上标注的收件人看到信的内容。隐私权倡导者认为，如果用户注册谷歌邮箱，就会把最隐秘的交流内容“泄露”给谷歌。现在，谷歌邮箱是全世界用户最多的电子邮箱服务，每个月的活跃用户超过10亿。他们大都清楚，作为交换，提供免费电邮服务的谷歌将“阅读”他们的通信内容。在知情、自愿的前提下，用户接受了这笔交易（包括个性化广告的投放）。

作为志存高远的一家企业，谷歌不会止步于搜索和广告这两项业务。2013年3月发布的谷歌眼镜原型机可以通过内置传感器，从佩戴者的视角观察、记录周围环境。批评者担心，谷歌公司有可能利用谷歌眼镜，在没有获准的情况下堂而皇之地分享人们的谈话内容。不过，可用于这个目的的技术似乎不只是谷歌眼镜，录音装置和微型摄像机也可以轻松地完成这项工作。手机中就有这种传感器，你是不是应该扔掉手机呢？我们不仅不会扔掉那些可以记录精彩时刻的设备，还会与谷歌眼镜等可佩戴技术形影不离。

在谷歌推出谷歌邮箱服务的几个月前，一个名叫Facemash的小型网站出现在哈佛大学的校园里。传奇人物马克·扎克伯格当时还是一名哈佛大学的学生。他编写了一个软件，利用在线电话簿，把9栋宿舍学生的头像“拖拽”到网上，让学生们从随机配对的两张照片中投票选出“更性感的”照片。这个网站深受扎克伯格同班同学的欢迎，也引起了极大的争

议。哈佛大学指责扎克伯格未经允许发表照片的行为侵害了同学们的版权和隐私权，这使扎克伯格陷入了困境。照片在新型交流形式中的应用，再一次满足了人类搬弄是非的欲望。布兰代斯法官肯定会感到惊恐莫名，但是不到10年时间，全世界已经有大量人口把脸谱网作为默认的交流工具。截至2016年，全世界有1/4的人口在使用脸谱网，超过10亿人每个月都会通过手机登录这个系统。扎克伯格正在不断突破社会规范的束缚，在他身后，一大批人迫切希望进入数字身份这个未知领域。

在发展壮大的过程中，脸谱网很自然地意识到，它们也可以像谷歌一样从事在线广告业务。将脸谱网上贴出来的内容用于定向广告，其潜力甚至超过邮件。人们在脸谱网上表明自己的情感状况、教育水平、政治信仰与宗教信仰等，列出喜欢的电影、电视节目、书和音乐，播报旅行情况，分享对一系列品牌和广告促销活动的看法。他们不断地上传自己、孩子、爱犬和宠物猫的照片。所有这些活动的本意是向亲朋好友提供关于自身的“公开”信息。2008年的一个夏日，我来到了脸谱网的总部。当时，脸谱网刚刚开始个性化广告业务，新广告有一个反馈意见按钮。如果某个用户不喜欢脸谱网向她展示的某个广告，并且点击了这个按钮，她就会收到阐述原因的请求。来自用户的这些真实的反映，令脸谱网备受启发。大部分人并没有抱怨这些广告过分地利用了他们在脸谱网上分享的个人信息，相反，他们认为广告对这些信息的利用不足。举一个典型的例子，“我的个人资料写得非常明确，我是男性，但我只对男性感兴趣。可是发给我的广告为什么还让我‘认识50岁出头的女性’呢”？用户要求他们看到的广告，应该与他们的真实需求有关。

2016年，脸谱网过了它的第13个生日。很快，我们就可以通过脸谱网亲眼看到一代人完整的童年生活了。在这些孩子还不能正式管理自己的脸谱网账户时，他们的父母和祖父母早已开始把这些孩子的生活情况分享到脸谱网上了。过去的中学毕业生有一堆身份文件：出生证、免疫证明、成绩单和毕业证书。大多数人还有驾驶证，有的人也可能有公司或宗教权威给他写的推荐信，或者护照。与之相比，出于父母、祖父母、叔叔阿姨、哥哥姐姐以及亲友们的原因，现在的青春期前儿童拥有了各种社交数据。你可以看到胎儿的超声波照片、抱怨学步儿童不愿遵守纪律的评论、生病时家人的祈祷，以及关于相貌、技能和嗜好的详细信息。脸谱网为什么要求注册账户的用户必须年满13岁呢？给每个刚出生的孩子建立脸谱网账户似乎更合理，这样就可以确保每个人都有一个独特、可靠的标识符，而且他可以选择用或者不用这个标识符。此外，

社交数据还可以贴上账户ID（账号）这个标签，在孩子长大以后（当社会规则认为他们有生活自理能力时），就可以管理这个账户ID上所附着的数据了。

过去，当我们置身于公开的环境之中，几乎没有体验过，也不指望享有隐私权。现在，一切都不同了，关于个人隐私和政治隐私的“权利”被放到卧室和投票室的墙上，虔诚地供了起来。随着互联网深入到社会存在的基本结构之中，为了免费、及时地联系亲朋好友以及远方的陌生人，我们甘愿将自己的生活“公之于众”。隐私权的逐渐形成与彻底取消，只花了200年的时间，不过是浩瀚的历史长河中泛起的一个浪花。

乡村野话

无隐私可言

烟囱和城市移民（17世纪初）

匿名社交与隐私权概念的提出

美国宪法第四修正案（1972）与无记名选票的启用（1856~1896）

隐私权登上政治舞台

“隐私权”（1890）

隐私受到法律的严格保护

谷歌、脸谱网及其他（今天）

隐私权是一种幻象——我们乐于分享

100年来，我们一直珍视隐私权，但是现在，我们必须认识到隐私权其实只是一个幻象。我们需要工具对我们关注的内容、亲密关系和交流沟通加以管理。布兰代斯法官提出了一个伟大的想法，但是这个想法仅属于他的那个时代——数据稀缺、社区本地化、交流成本居高不下的

时代。在那个时代，如果你愿意，你可以轻松地阻止他人发布你的照片。但是，这在今天根本办不到。此外，匿名性并不是民主的默认设定。更明智的做法是针对当下的现实和未来的前景制定规则，而不是把隐私权理想化，奢望过去的规则未来仍然可以保护我们。为了让数据为人类造福，我们需要透明性和主动性。

我们不应花费精力去考虑哪些数据可以公开，哪些数据必须保密，然后不厌其烦地对它们区别对待，而应该全神贯注，尽可能地不矫揉造作。只有这样，我们才能充分利用数据服务商的产品，使数据分享可能带来的积极影响和消极影响取得平衡。

在互联网上，所有人都知道你是谁

社交数据，再也不涉及你是否享有隐私权的问题。1993年，彼得·施泰纳（Peter　Steiner）在《纽约客》杂志上发表了一幅经典漫画，漫画的题目非常有意思："在互联网上，没人知道你是坐在电脑前的一条狗"。不过，情况已经今非昔比了。适用于今天的漫画题目应该是："在互联网上，所有人都知道你是一条狗。你戴着蓝色的颈圈，对猫感兴趣。你的主人正在度假"。这是因为你把这些信息分享给了数据服务商，以便与好友交流和收到个性化的商品推荐。你需要为此付出的代价之一是收看康多乐狗粮广告。因此，在互联网上隐姓埋名，不过是人们的一个臆想。

但是，早在脸谱网问世之前，数据就可以暴露个人身份了。20世纪90年代中期，计算机科学家拉坦亚·斯威尼（Latanya　Sweeney）决定研究某个"无记名"医疗数据库的匿名性。马萨诸塞州政府认为，将州政府雇员的医疗就诊信息分享给研究团体的做法更符合公众利益。州政府官员并不傻，他们知道连名带姓地分享这些数据是不合适的，因此他们去掉了身份标识，例如每个人的姓名、住址和社保号。但是，为了有利于改善医疗政策，他们保留了几个相关数据，包括性别、出生日期和邮编。斯威尼将这三组数据和另一个数据库（剑桥市选民登记册。只需缴纳20美元，即可公开使用该数据库）的数据进行了比较，结果她准确地找到了马萨诸塞州州长的医疗记录。"最夸张的是，斯威尼博士将这份医疗记录（包括医生的诊断意见和处方）寄到了州长办公室。"

斯威尼估计，如果知道性别、出生日期和邮编，就可以确定87%的美国民众的身份。后期研究把这个数字缩小到接近63%，不过，考虑到无须了解更独特的个人特征（诸如人们每天在脸谱网等社交数据网站上分享的个人身份信息），这个数字仍然高得令人吃惊。为什么只需要如此少的数据即可精准确定一个人的身份呢？只需粗略计算，即可发现其中的奥秘。美国正在使用的邮编约有4万个，总人口约为3亿，平均7 000人使用同一个邮编，其中男女大约各占1/2。假设一个日历年中每一天出生的人数相同，那么在每个邮编对应的人口中，生日为同一天的人数就是这一年出生人口的一个因数，大约为10名男性和10名女性。

现在，我们考虑数据服务商通常可以获取哪些社交数据。如果两个大型数据服务商同研究人员分享“不记名”的社交数据，认为无法通过数字痕迹确定人物身份的想法就会被彻底击碎。第一个数据服务商是美国在线（AOL）。AOL曾经公开658 000名用户在三个月内的不记名搜索日志，供学术研究之用。但是，由于公开方式选择不当，任何人都可以下载这些数据。《纽约时报》的两名记者根据这些搜索记录，成功地确定了几个人的具体身份。他们之所以轻松得手，是因为人们经常搜索自己和亲友的信息，以及从家出发前往某地的路线。第二个大型数据服务商是网飞（Netflix）。网飞为了更准确地预测用户对电影的评分，宣布举办一个竞赛。因为参加竞赛的研究人员需要相关数据建立模型，网飞就提供了48万名用户的“1亿个电影评分结果，以及完成这些评分的日期”，其中不包含用户的姓名。但是，得克萨斯大学的两名研究人员阿尔温德·纳拉杨（Arvind Narayan）和维塔利·施玛蒂科夫（Vitaly Shmatikov），通过比较IMDb.com网站公布的电影评论与不记名数据，对网飞的用户数据成功地实行了“去匿名化”。既然那些电影评论都已经是公开信息，那么网飞数据的匿名化又有什么意义呢？无名氏诉网飞案的原告认为，网飞的用户不会每租一部电影就发表评论，他们挑选的某些“私密”电影会暴露过多他们的个人信息。由于有5万名研究人员获准使用网飞的数据库，这位原告担心她的同性恋身份已经暴露了。（之前，为无名氏提交诉状的那位律师曾迫使脸谱网停止提供将用户租借热门录像带的信息分享给他的好友的特别服务，用户只能通过“选择性加入”按钮才能分享这类信息。）

即使你对公开电影租借记录这件事感到无所谓，但如果将你的搜索记录完整地公之于众，你也可能感到很不舒服。只要不是异于常人，你在谷歌地图中输入最多的地址就应该是你的家庭住址了。你住在哪里，

喜欢去哪些地方，需要买什么，你经常了解哪些人的信息，你为什么对他们牵肠挂肚，这些问题都涉及最隐秘的生活细节。搜索项可以反映人们关注的问题，谷歌通过Google Trends（谷歌趋势）将人们关注的问题展现在我们眼前。很多人看到新闻报道网络化的趋势，Google Trends还表明，在过去两年里，人们对“网络暴力”、“跨性别者”的兴趣在不断增加，而搜索“隐私权”和“变性人”的人数则呈下降趋势。

现在，我们设想一下，如果能实时看到个人的搜索行为，会怎么样呢？20世纪90年代，我去斯坦福拜访在一家互联网搜索引擎类初创公司就职的朋友，看到了一些正在进行中的搜索行为。其中一个人的搜索项是“如何自杀”，这引起了我的注意。如果换成是你，你会怎么办呢？你会根据他的互联网服务商和IP（网络之间互连的协议）地址追踪这位用户，然后拨打自杀求助热线吗？这样做，会不会侵犯用户的个人隐私权呢？为了弄清楚用户这次搜索行为的确切含义，正确解读他的动机，更具体地确定他的自杀行为发生概率，你会不会首先查询这名用户的相关搜索记录呢？也许他是一名小说家，搜索“如何自杀”是为小说创作收集资料，而不是打算自我伤害。但是，接下来你又看到这名用户正在搜索“金门大桥”，有超过1 600人在这里结束了自己的生命。在这种情况下，你的注意力会不会离开显示器，离开那名用户，深吸一口气，接着考虑改进搜索结果质量的问题，而将这名随时可能结束自己生命的人抛在脑后呢？这个问题可不是那么容易回答的。

与之类似，电子商务也会暴露你的一些特点，有时还会连累他人受“池鱼之灾”。想要亚马逊送货上门的话，你就必须提供个人信息，包括姓名、送货地址等。提供正确的地址对你有利，否则你就无法收到包裹。不过，购买记录既包含你买给自己的商品，还有你买给他人的商品。如果你将某件商品标记为礼品，亚马逊在为你做产品推荐时就不会考虑这件商品。利用这些数据，个性化算法在处理你备注为他人购买的商品时，就会将它与你购买的其他商品区分开。如果你为某位女性买礼物，那么你在选择衬衫尺寸时就会公开她的体型数据。如果你买这件衬衫的时间是在母亲节之前的一两周，而且收件人的姓氏与你相同，亚马逊的算法就会推断出你们之间的关系。一年后，亚马逊甚至有可能给你发电子邮件，推荐适合你的母亲节礼物。

“你的亚马逊”页面为用户提供了一定程度的透明性和主动性。在这里，你可以看到自己的一些原始数据，包括购买记录，还可以决定哪些

数据可以用于个性化商品推荐。你也可以把你在其他地方购买的物品添加进来，无论是你近期购买的还是几十年前购买的。2014年，脸谱网采用了类似方式，允许用户进入“活动日志”页面，其中列出了一系列好友请求、点赞、你身在其中并被标签的故事与照片、需要回复的事件以及其他内容。如果你愿意，还可以从记录中删除某些数据点。由于脸谱网上的数字身份可被用来生成个性化广告，因此删除脸谱网历史记录中的某些内容有助于你对发送给你的广告内容施加影响。

即使从历史记录中删除一两个，甚至20个点赞行为，也不大可能隐藏你行为举止的总体模式。事实上，剑桥大学心理测量中心的戴维·史迪威（David Stillwell）在研究中发现，脸谱网活动日志可以相当准确地反映出用户的个性特点。史迪威招募了数千名脸谱网用户参加测试，评估他们在五大个性品质方面（开放性、尽责性、外向性、亲和性和情绪不稳定性）的特点，他还请另一组实验对象认真研究这些用户的脸谱网档案，并对他们的个性做出评估。结果，两组实验对象评估结果的吻合程度高到令人吃惊。人们习惯在脸谱网上准确地展示自我——他们不喜欢伪装，在管理社交媒体上的个人档案时也如此。如果一群陌生人可以根据你的脸谱网“个人时间轴”准确评估你的个性，那么算法也一定能做到。在脸谱网上，你为好友及时更新你的状况，但这是需要付出代价的，会暴露你在尽责性方面有所不足的缺点。

2013年，史迪威与同事迈克尔·科辛斯基（Michal Kosinski）以及微软研究院的一个团队合作，推出了YouAreWhatYouLike应用程序，评估根据用户在脸谱网上的行为推测其个性品质（包括智商、种族、政治信仰、易成瘾物质使用情况、性取向等）的效果。研究人员称，在88%的案例中，他们的“模型可以准确地区分同性恋与异性恋男性”，依据仅是脸谱网上的点赞行为，尽管这些行为与任何政治问题或政治权利都没有明显的关系。研究人员说，对辨识男同性恋者有明显帮助的是他给“魅可化妆品”“坏女巫音乐剧”等点赞，而给“嘻哈乐队武当派”和“小睡刚醒有点儿迷糊”点赞的大多是男异性恋者。企业被允许利用智商测试和人格测试来筛选求职者。也许在不久的将来，有人会要求你安装一款类似的程序，以测试你自我评价的高度条理性、处变不惊是不是自我吹嘘。

即使你不做出积极贡献，能反映你的个性特点的数据也可能集腋成裘。网上照片泛滥成灾就是一个明显的例子，你的所有照片并不都在你的掌控之中，更不用说拥有版权了。如果你参加会议，或者别人拍照时

你恰好从镜头前经过，那么你的脸孔被人认出只是时间早晚的问题。雅恩·乐昆（Yann LeCun）领导下的脸谱网小组开发的人工智能程序可以识别两张照片上的人是不是同一个人，而且识别的准确程度已经与人类相差无几了。这个名叫“DeepFace”的智能系统还不能根据脸部照片确定姓名，但是，如果两张照片上的人似乎是同一个，而且有人用某个姓名来标签其中一张照片，那么算法也会把这个标签赋予另一张照片。研究小组正在开发其他软件，用于分析照片背景和具体环境，以判断你是置身于熙熙攘攘的酒吧还是待在荒无人烟的山顶。如果你出现在其中一个场合中的次数更多，算法就可能把你归到社交蝴蝶或者孤胆探险家一类。

微软研究院的辛西娅·德沃克（Cynthia Dwork）等人已经证明，正因为有了数据和数据库，人们才有接触信息的机会。数据库的目的是答疑解惑，如果有一系列疑问，数据库中可能只有一个人能够正确地回答全部问题。辛西娅经常举例证明这个事实。她先问微软员工医疗数据库中有镰状细胞特征的人占多少比例，然后问除了受人尊重的卷发女科学家以外，有多少人有这种特征。由于辛西娅是微软员工中唯一一名受人尊重的卷发女科学家，因此这两个答案之间的差值就能告诉你她是否具有这种特征。

人们把数据分享给数据服务商，目的是得到有助于决策活动的个性化精炼数据。辛西娅·德沃克描述的那种数据库收集的数据种类比较具体，受到的限制较多，也就是说，收集的是“小数据”。相比较而言，现代数据服务商收集的数字痕迹都是令人难以置信的“大数据”。为了从数据服务商那里得到有用的产出，你必须愿意提供准确的数据输入，例如你真正的兴趣和喜好。如果你不愿意分享这些数据，你收到的商品推荐信息不会比普通人好，也就是说，你会收到最受大众欢迎或者与他们关系最紧密的商品推荐信息。如果你提供的数据不正确，那么你得到的产出对你来说可能毫无用处。如果你希望享有多一点儿的隐私权，就要付出效用降低的代价。

使用假名的利与弊

同意或拒绝提供身份辨识信息的决定会造成某些后果。暴露身份到底会不会导致某种风险或危害，需要视具体情况而定。数字痕迹清晰可

辨，使我们几乎无可遁形。

不过，直到脸谱网面世之后，用户们在社交平台上使用真名才成为一个常见现象。此前，使用假名是习惯性做法。在一定程度上讲，这是逻辑学导致的问题。有的人的姓名十分普通，因此，在没有其他手段区分用户的时候，不可能让所有用户的账户名都与他们的真名一致。此外，有的网站没有为账户名留出足够多的字母位，以至于较长的姓名不能用作账户名。在有的场合下，人们也不希望暴露自己的真实姓名，担心身份被盗用、被追踪或因为发表不受欢迎的意见而给工作或生活带来不良影响。在任何情况下，如果你愿意，你就可以在你使用的社交媒体平台上给自己取一个甚至若干个不同的账户名。结果，在互联网问世后的头几十年里，一个显著的标志就是数量空前的不完整身份。在这个过程中，我们摸索出同他人交流的新方法。

人们的传统身份包含一些简单数据，例如姓名、出生日期、身高、眼睛颜色、国籍、居住地等。通过这些基本信息，可以确定你真的是你。很多规章制度在实施时都需要验证人们的身份。几百年来，我们一直利用身份证来证明我们可以合法进入某个地方，利用支票和支票保证卡来证明我们存储在银行金库里的现金足以支付某件商品的费用。你的年龄或者公民身份允许你享有某些社会权利或者担负某些社会义务，例如投票、在公共场所饮酒的权利或者缴税、服兵役的义务。我们逐渐接受了一个事实：在生活中，我们经常需要拿出政府发放的身份证或报出身份证号码，输入密码，或者回答一系列问题，例如我们的常飞旅客账号或童年时期养的宠物是什么。

很多数字痕迹都是你与物理设备交互时留下来的，而且有为数不少的交互过程非常独特，足以表明你的身份。由于人们利用手机和平板电脑接入网络的时间在不断增加，许多数据服务商投入了大量研究资源，试图根据用户行为举止的规律性，综合各个设备的数据，确定某个人的身份。他们使用的一个方法是要求用户登记注册，但如果同一个人使用不止一种设备，就会留下更多的线索，例如他选择的字体。此外，某些人会不断犯同样的键入错误，他们发现并纠正这种错误的频率高于他人，这个规律也可用于探查用户的身份。

身体接触设备，也会留下痕迹。以色列网络安全初创公司BioCatch的联合创始人尤里·里夫纳（Uri Rivner）认为，数字指纹可以表明用户操作电脑、平板电脑或手机的手法，“通过观察你的行为与行为方式来验

证你的身份”。为收集数据，BioCatch公司要求用户完成某些可以验证其身份的操作，但是不告诉他们这样做的目的。公司感兴趣的不是你搜索的内容，而是你搜索这些内容时所采用的方法。使用触摸屏时，你是用力按下还是轻轻地抚摸？拿手机时，你的手颤抖得厉害吗？上下滚动屏幕时，你会点击屏幕的哪个部位？你移动鼠标的速度有多快？你喜欢打开新标签页还是在现有标签页上点击前进或后退项？BioCatch公司的客户包括一些需要用新方法验证顾客身份的银行。

实时数据分析还可以在其他领域用于身份确认，例如，在某些领域里，资格证书不足为信或者没有资格证书。面向儿童的网络、应用程序或游戏必须考虑用户是否安全、内容是否适合等一系列问题。为6~16岁用户提供游戏服务时，网站至少要确保推荐给每名用户的都是适合他们的游戏。游戏开发商需要考虑的问题不只是8岁儿童是否愿意尝试那些评级为仅适合10多岁青少年的游戏。如果游戏设置太难，孩子很快就会放弃，但如果设置太简单，孩子又会觉得没意思。由于某个家庭成员登录游戏网站之后，兄弟姐妹很可能使用同一台电脑玩游戏，因此网站不能完全相信用户资料中的年龄数据，而是通过分析游戏者与游戏之间的互动，来估计他们的年龄。网站经常使用的一个安全措施是规定游戏者在聊天对话时只能选用预设语句，目的是降低游戏者面临的风险，以免他们在不经意间将家庭住址等关键信息透露给伪装成青少年的成年人。研究表明，年龄稍大的孩子所选择的预设语句通常不同于他们的弟弟妹妹。此外，游戏网站据说可以根据鼠标移动特点推测孩子的年龄，而且误差仅为3~6个月。对于10岁以下或刚刚10岁的孩子而言，他们的鼠标移动特点与其运动技能的发展情况密切相关。

想要愚弄某个寻找这些不明显痕迹的机器学习系统，难度比假冒某些明显特性要高得多。在医院里，如果一位穿白大褂、脖子上挂着听诊器的人要求你脱下衣服，你也许会认为这个人是一位合法的医生。不过，人们已经发现，出于这样或那样的原因，有些人会假冒某种身份。2015年1月，一名17岁的年轻人身穿白大褂、挂着听诊器，骗过佛罗里达州的一个医学中心的保安，在医院里冒充医生。直到一个月之后，他才被拆穿，被警察拘禁。

在历史上，人们曾经利用假名来实现言论自由。1787年，刚刚颁布的美国宪法受到了人们的严厉批评，为此“普布利乌斯”［其实是亚历山大·汉密尔顿（Alexander Hamilton）、约翰·杰伊（John Jay）和詹姆斯

·麦迪逊（James Madison）三人］发表了一批论文，对这些批评进行了反击。在辩论过程中，几乎没有任何人发现他们的身份。乔治·艾略特（George Eliot）出生时名叫玛丽·安·伊万斯。她在一篇文章（一如既往地使用了假名）中说，19世纪的女作家只会写一些“愚蠢的小说”，其特点是“空洞、单调、虚伪或者迂腐”。为了摆脱人们普遍持有的这种偏见，她开始使用笔名。艾略特希望人们认真对待她笔下的人物和她的文字，但她认为，如果读者因为封面上的作者姓名而对作品持有偏见，她的努力就会化为泡影。

有时，使用假名并不是出于追求言论自由这个动机，而是希望与既往的历史决裂。1947年，一位名叫汉斯·法拉达（Hans Fallada）的人（他的“真名”是鲁道夫·迪岑）出版了小说《每个人都孤独地死去》，讲述一对德国夫妇秘密反抗纳粹的故事。法拉达受到一位苏联文化专员的委托，在研究国家秘密警察的卷宗之后，写就了这本伟大的反法西斯小说。不过，法拉达这个假名的诞生时间比这本小说早好几年，因此他并不是担心他的优秀小说家身份会与这本政治敏感的小说联系在一起，而是希望将他的写作与他在文坛的名声同令人难堪的自杀企图割裂开。

这三个著名的假名有一个共同点：主人都希望长期使用这个假名，并且赢得了声誉。“普布利乌斯”的所有作品都在为美国宪法获得批准摇旗呐喊，艾略特与法拉达出版的所有作品都使用了笔名。这些作者都希望他们的独创性的产品与一个单一身份联系在一起。

互联网刚刚问世时，有多个假名似乎是一个非常好的选择。遗憾的是，多个假名会导致一个问题。想出一个新假名并不难，但你怎么知道这个新账户名与一周前被网站拉黑的那个家伙没有任何关系呢？网站可能坚持要求注册账户时使用假名必须登记电子邮箱地址，但是创建电邮账户同样容易。有的平台要求用户在注册时填写非常复杂的申请表，这项举措会略微增加新账户的创建成本，但无法阻挡“有献身精神”的骗子，他们可以雇用大量人员或利用机器人来填写这些表格。“廉价假名的社会成本”［经济学家埃里克·弗里德曼（Eric Friedman）与信息科学家保罗·雷斯尼克（Paul Resnick）杜撰的名词］是无法通过这种方法消除的。

能不能将假名的成本增加到足够高，使它们与真名一样可靠呢？这得视情况而定。如果信任需要从第一次交互时开始培养，那就应该使用“真名”，因为“真名”可以展示用户行为的历史记录，比如，你与银行

或信用卡公司打交道的记录。与之相反，使用假名则需要从零开始，慢慢建立自己的声望。

我在亚马逊就职的时候，研究过这样一个问题：利用假名与利用真名发表的顾客评论，哪一种对其他用户的价值更大？我们知道，登录亚马逊账户，并使用某种类型的假名，就会降低发表“无用”评论的可能性。我们还知道，顾客更重视非匿名评论。只要亚马逊的顾客改变自己的账户名，他的所有评论（包括过去和现在的评论）的署名就会更新为这个新的名字，使每个人的评论历史记录保持不变。顾客的身份与评论历史记录都具有持续性，而展示给公众的假名则未必如此。亚马逊本来可以要求评论者使用真名，因为亚马逊的每一个顾客都是有真实姓名的，这一点已经通过与账户绑定的信用卡得到证实。不过，他们发现，最重要的因素是亚马逊能否判断该评论者确实购买了那件产品。人们的确更信任使用真名的用户的看法，但就亚马逊的顾客评论而言，人们更关注的是评论者购买商品的数据，而不是他的姓名。根据这个发现，亚马逊改变了产品“平均”评级的计算方式，为真正购买的顾客评论赋予了更大的权重。

使用假名还需要做出其他妥协。想一想，在经常光顾的地方填写纸质“评价卡”与接受网上调查之间有哪些微妙的区别？尽管从表面上看，评价卡是不记名的，但是很多人不愿意填写，原因不是因为懒惰，而是他们知道可能会因为若干特性暴露自己的身份，包括笔迹、遣词造句、提出的主题，以及他们将评价卡放进回收箱的时间等。他们担心，分享负面评论可能会造成不良后果。顾名思义，不记名评论也是“一次性博弈”。双方之间没有对话，没有澄清语义和解释意图的机会，商家也没有为顾客的配合提供奖励。因此，商家不仅不重视这类反馈意见，而且将其视为噪声、异常信号，认为这些评论仅适用于当前情况，并不表示需要做出改变。匿名性可能导致反馈意见被视为有私心或者恶意行为，并因此遭到无视。

红迪网等在线讨论平台必须通过机器学习解决有关匿名性的这些问题。红迪网的账户名既可用于用户同整个社区的每一次交互，也可只用于某一个帖子或某一次投票。每一个假名都有畅所欲言的自由，网站还鼓励在平台上发表评论的人尝试使用各种各样的形象。他们从来不要求用户为账户绑定电子邮箱或者填写真实姓名，网站创始人不希望通过这种方式管住人们的嘴，责任心需要用其他方式培养。你发表有意思的言

论之后，其他人就会参与进来附和你或者反驳你，发表他们的评论，或者为其他人的评论点“赞”或“踩”。

如果某个帖子或者评论经常被“踩”，就会显示为灰色，被放到排序列表的底部，并附上“评论分低于阈值”的说明，但人们仍然可以点击浏览它，也可以对它进行评论。网站没有把用户的看法束之高阁，而是允许用户展开对话，自行判断哪些评论值得一看，哪些评论则纯属无稽之谈。

对红迪网来说，最重要的是那些荣登“热门”、“排名上升”和“有争议”名单的讨论，因为这些讨论真的可以吸引很多人，而不是因为可以使用假名。在任何一个名单上跻身前25名的讨论常常会在互联网上引起广泛关注。红迪网没有花费大量的时间与财力，安排“人类仲裁者”制定实施各种规章制度，而是依靠机器学习减少“投票骗子”（这些人使用不同的用户名，但他们的目的不是发表意见，而是为自己的帖子点“赞”、给其他人点“踩”）的数量。如果多个假名同时处于活跃状态，而且使用同一个IP地址或者写作风格相似，这些假名就会被视为合谋者“圈”到一起。合谋者的投票得到的权重较小，偶尔还会被忽略不计。

真实的信号

2016年，为了与人邂逅、约会、谈恋爱和建立长期关系，有超过1亿人使用了相关的应用程序和网站。如何才能找到满足彼此要求，并且兴趣相投（这个要求最难实现）的人呢？

说到约会，有的人在某些方面表现得比较诚实。诚实度的分布状态因人而异、因具体情况而异。有时，人们也许只是做个实验，自己到底想要什么。人们的言语是一种信号，行为是另外一种信号。通过实际行为传递的信号就是社会科学家所谓的“真实信号”。

交友程序的用户界面和推荐算法的设计难度极大，因为用户可能自认为中意有某种个性特征的人，但是他们使用交友网站的情况却表明他们的心仪对象是另一种个性特征的人。OkCupid在线约会社交网的联合创始人克里斯蒂安·拉德（Christian　Rudder）发现，用户可能没有完全意识到或者根本不想承认，种族偏好和民族偏好对他们择偶的影响程

度。但是，对点击数和联系人信息稍加统计，就能迅速揭示这些偏好的存在。

这与电影评分这个老问题有点儿相似。网飞邀请用户评价《公民凯恩》等特别受欢迎的电影或者纪录片《黑鲸》，很多人都打了五星，因为他们认为既然“所有人”都说好，他们也应该给高分。网飞可以根据你的评分给你推荐电影，但这些评分必须诚实可靠才行。此外，必须有证据表明诚实评分对用户有利，而且要让用户看到这些证据。网飞发现，能够更好地反映人们对哪一类电影感兴趣的真实信号，是他们在线观看这部电影的实际时长。换言之，对于形成推荐意见而言，检视数据比评论数据更有帮助。这个现象与理查德·尼斯贝特（Richard Nisbett）的一个发现有关系：人们常常不理解行为与决定背后的认知过程，我们的自我了解与自省能力都是有限的。

在现实世界中，某些偏好和特性似乎很不明确。如果一位很棒的潜在约会对象比理想对象的年龄大了几个月，大多数人都不会在乎。但是，我在几家交友网站担任咨询师的时候，发现说自己29岁比说自己已满30岁的人要多得多。这个现象与事实不相符。这些年龄造假者的欺骗行为是发生在他们建立个人档案的时候，还是在与应用程序交互之后发现自己“太老”了——当他们感兴趣的约会对象在网站上搜索时，他们的资料不会出现在其屏幕上，所以撒谎说自己29岁呢？这个现象引发了我的思考：如果用户编辑个人资料的历史记录可见，这有没有可能改变他们的行为呢？

有的个人资料编辑行为是可以接受和理解的。例如，某个人与若干对象约会后，决定修改他在个人资料中列出的兴趣爱好，因为他觉得夸大了自己的攀岩技术，或者低估了自己对音乐会的热衷程度。同样，他也可能修改对心仪对象的描述。但是，有的修改可能令其他用户不快，例如，一会儿说自己单身，一会儿又说自己正在与某人热恋，而且变化频率非常快。

假设用户们不仅可以看到彼此的个人资料编辑记录，还可以看到交流记录，会怎么样？异性交友应用程序经常会遇到一个问题：女性用户接收到的信息泛滥成灾，而男性用户却无人问津。为了平衡这个态势，交友程序必须限制用户在某个时间段里发送信息的次数。但是，由于交友程序的用户每天都在变化，人来人往，因此限制信息发送次数的方法注定会失败。如果你用完了每月的信息发送次数，而这时你的“白雪公

主”出现了，等到你可以再次发送信息的时候，她早已消失不见了，你该怎么办呢？算法把她的资料淹没在搜索结果之中，是因为你没有及时与她联系，还是因为她已经与他人约会了呢？你不得而知！因此，硬性限制发送信息次数的办法并不可取，更好的做法是借助透明性原则，发现用户行为的真实信号。例如，每个用户的个人档案可以显示该用户前一天、前一周和前一个月里发送和接收信息的数量，以及平均回复率和反应时间。这些信息将有助于你判断哪些人可以接触。

某些交友应用程序已经启用了这种面板数据。男同性恋交友应用程序Jack’d会提供每个用户的回复率，并通过一些描述性数据（年龄、种族、体型等的分布情况）给用户推荐其真正感兴趣的目标对象（不只是用户在个人资料里声称的目标对象）。这种透明性不仅有利于用户充分考虑所有的选择，还不会让他们错失良缘。如果你希望接触的人的回复率只有12%，那么你也许宁愿去接触其他人。如果引起他注意的用户中有64%的人声称自己有“大块肌肉”，而这又恰恰是你的短板，你就更不需要浪费时间了。简单地说，在提供用户品位数据时，Jack’d的依据并不是点击数、收信数或发信数，而是用户的“喜爱类型”清单和“电子红娘”交友工具。“电子红娘”允许用户表达自己的兴趣，但只在双方都互相表示有兴趣时才会发出提示信息。

除了这些发给其他人的明确信号之外，所有交友网站还存储了大量更加隐秘的信号，即每个用户在浏览其他用户的个人资料时留下的痕迹。但是，这些浏览行为背后的动机非常复杂，难以理解。我在Match婚恋网工作期间，发现一个用户屏蔽了大量黑人女性。据此最容易得出的结论就是这个用户是种族主义者，对吗？错！我们研究了他的过滤器设置和浏览记录，结果发现真实情况正好相反。他只对黑人女性感兴趣，特别是那些自称“曲线优美”的黑人女性。他曾向一些黑人女性递出橄榄枝，但遭到了婉拒。为了不浪费自己的时间和精力，他屏蔽了这些女性。作为一名数据侦探，解决这样的疑难问题是一大乐事。构思好的故事，并讲好这些故事，是数据解读的必备技能。

用数据讲故事，你必须想办法了解用户的想法。同所有故事一样，情境是一个重要因素。为了实现这个目标，我们经常忙得昏天黑地。我曾为交友网站Fridae出谋划策，它是一家位于新加坡的网站。当时，我们发现，用户在周五下午2点和周日凌晨2点浏览的其他用户的个人资料是不同的。因此，Fridae交友网站的数据科学家团队必须考虑，如何据

此对推荐给用户的约会对象排序。

越来越多的交友网站鼓励用户添加脸谱网个人资料或者Instagram、推特账户的链接，以展现真实自我。但是，这并不意味着恶意欺骗行为已经消失殆尽。手机交友平台Skout（交友乐园）的数据科学家塞巴斯蒂安·波尔（Sebastiaan Boer）建立了一个过滤不当信息的算法，并给它起了一个好玩的名字——“讨厌鬼终结者”。哪些信息是不当信息呢？它是指依据用户点击与互动记录被判定为令人讨厌的信息。如果某人遭到多名用户的屏蔽，他就可能是一个令人讨厌的家伙。如果某人单方面给某个用户多次发送信息，那么在后者的心中，前者就非常讨厌。算法逐渐学会判断，涉及哪些内容的信息会导致用户屏蔽该信息或者得不到回复。悄然出现的否定性信息就是一种典型的不当信息。“讨厌”、“丑陋”等词语是一个标志，但是不当信息的定义远非如此简单，因为这是一个仁者见仁、智者见智的问题。在信息屏蔽表现出某种规律性之后，“讨厌鬼终结者”就会阻止不当信息的发送。此外，如果针对某个用户发送大量信息，却从未得到任何回应，这种行为就会被阻止。“讨厌鬼终结者”的目标是为大多数用户提供一个积极的交友环境。

我在本章开头告诉大家，学习物理学的那段经历和物理学知识对我利用社交数据设计、完成和分析实验有所帮助。许多社交数据实验需要观察数据挖掘程序设计的变化是否会影响人的行为。如果交友程序用户看到他感兴趣的对象很少回复信息，那么他会字斟句酌地编写信息，以期引起对方的注意，还是会转而寻找更有可能回复他的交友对象呢？在哪种情况下，一个令人讨厌的家伙更可能停止发送不当信息的行为，是遭到系统管理员的阻止，还是从未收到任何回复？试验身份特性的行为是否会改变用户的回复率？在试验身份的过程中，一旦超过某个底线就会被其他用户视为骗子，这个底线是如何确定的？增加用户行为的透明性，有利于用户自行判断某个用户个人资料反映出的性格特征，是否与其理想的灵魂伴侣的特点相吻合。

隐私权和责任心不可兼得

说到隐私权和责任心，人们总想自己享有隐私权，而要求其他人有责任心。

——戴维·布林（David Brin）

由于人们利用移动设备接入互联网的时间不断增加，另一种假名——电话号码，逐渐变成了身份信息的一个重要组成部分。第一代家庭电话投入使用时，接线员会通过电话告诉你另一方的身份，并询问你是否接听。随着转盘式脉冲拨号与自动交换技术的发展，人们可以直接通话。要想知道电话的另一端到底是谁，你就得暴露你方便接听电话这个秘密。不过，只要打电话的成本居高不下，每个家庭就不会接到过多不想接听的电话。电话费不断下降，在经济这个层面上为电话推销打开了方便之门。1990年，几乎在网络发明的同时，音频拨号系统问世了，这为来电显示（Caller ID）服务创造了条件。

最初，这项将电话号码（可能还包括你的姓名）自动传送给被呼叫者的技术，遭到了一定程度的抵制。但是现在，情况已经发生了改变。如果呼叫者不表明自己的身份，人们是不大可能接听他的电话的。“未知”号码的来电都被转接到语音信箱了。想要对方接听你的电话，你就得让他知道你是谁。也许你觉得比较安全的做法是不分享自己的电话号码，也许你还认为，如果呼叫者告诉你他们的号码，就更安全了。但是，只有建立在公平的基础之上，也就是说，彼此都知道对方的身份，交流的效果才会更好。

在线电话号码供应商白页公司的创始人亚历克斯·阿尔加德（Alex Algard）认为，强制要求增加电话交流的透明性的做法不可能对所有用户都有利。Hiya（原“白页呼叫者身份”）程序无视呼叫者的设置和被呼叫者手机上的通信录，直接对来电进行身份鉴别。在手机“垃圾信息”日益泛滥的今天，这项技术的价值更加凸显出来。通过挖掘在线资源，分析某个号码向Hiya程序用户呼叫的规律，Hiya程序可以将该号码归入某个类别（例如“电话推销商”）。作为机构，它们必须判断通话双方是否有权利知道对方的身份。如果有，它们还需要判断应该如何利用这些数据。

不过，这个问题有点儿复杂，因为持续性身份识别是产生信任的必要但不充分条件。知道某人是谁的意义并不大，只不过在他行为不当时，你可以要求他做出解释。我认识的一些人把他们在交友应用程序上看到的他们认为的不当行为进行了屏幕截图，并上传到脸谱网上。第一种情况是对方不接受“我不感兴趣”的回复；第二种情况是对方有侮辱性言论。两名收信人可能都点击了“屏蔽”按钮，然后就不再联系了，但他们都决定将这些不当行为的相关信息分享给好友。

今天，你在与人交流时，对隐私权抱有多大的期望呢？上面谈到的那两名收到不当信息骚扰的用户可能认为，分享自己的这些“隐私”信息有利于他们朋友圈中的好友。这些屏幕截图可以警告那些可能正在使用交友程序的人，让他们提防这些不当行为。的确，犯错者的照片、账户名都清清楚楚地出现在截图中，在收信人的脸谱网好友中，他们将无处藏身。同样，如果老板在电子邮件里不公平地指责你，你也可以将邮件转发给好友，或者发布到网上。法律可能认为你的这种做法不对，因为这封电子邮件涉及“机密”内容，不应散布到公司以外的地方。但是，分享这封电子邮件对公众有利，便于潜在员工加深对该公司工作环境的了解。

我们如何看待某人分享私密交流内容的做法，在一定程度上取决于我们对这个人的信任程度。在像红迪网这样的讨论平台上，用户基本上都是匿名的，你几乎没有办法鉴别发帖者的身份，更不用说辨别这些信息的真伪了。而脸谱网上的发帖者通常都是我们认识的人（或者发帖者与我们认识一个人），再加上账户被盗的概率极低，因此用户在决定贴出屏幕截图时心有顾虑，担心好友知道他将他们之间的隐私信息分享给其他人。不过，这并不意味着我们应该认为屏幕截图都是真实的，因为也有可能是某人为了败坏他人的名誉而故意伪造的。

一旦屏幕截图（无论真假）被上传，它与其他数据就没有多大区别了，任何人看到之后都有可能将它分享给其他人。如果脸谱网用户在收到不当信息之后勃然大怒或者感到好笑，并截图分享给他的好友，会怎么样呢？又或者，如果他的好友决定把这张屏幕截图发到推特上呢？接下来，肯定有人利用人脸识别算法去鉴别这张图片，然后标上发信者的姓名。至此，照片的初始背景（包括发帖人的身份）已经被剥离，但是，任何人搜索关于那个人的信息，都可以了解整个事件的来龙去脉。

未来，人们可以采取哪些措施来保护自己在网络上的名声呢？2014年5月，欧洲法院做出了一项有利于要求“被遗忘权”的判决，使我们多了一个选择。一位西班牙人在求职与租房时屡屡遭拒，因为雇主与业主都看到了一篇报道他由于没有缴纳税款，即将失去住房的文章。即使他后来补缴了税款，雇主和业主依然拿那篇文章说事，这令他十分烦恼。他没有要求清除他拖欠税款的官方记录，而是希望当人们在谷歌上搜索他的姓名时，不要让那篇文章出现在搜索结果中。法院认为，如果人们认为某些网页导致自己受到了伤害，那么他们有权要求将这些页面从搜

索结果中剔除。在欧洲法院的这项判决生效的第一天，谷歌就收到了超过1.2万个这样的请求。一年之内，一共有27.5万封请求信涌进谷歌公司。

谷歌公开了在这位西班牙人胜诉后收到的部分链接删除的请求。比如，一位意大利妇女要求将涉及10多年前，她的丈夫谋杀案的文章从她的姓名搜索结果中删除，一位在参加抗议活动时受伤的拉脱维亚活动家要求将涉及该次抗议活动的文章从他的姓名搜索结果中删除；一位10多年前“被判决犯有轻微罪行”的德国教师要求将涉及那次判决的一篇文章从自己的姓名搜索结果中删除。在这些案例中，谷歌认为个人的被遗忘权重于“公众对内容的兴趣”。这些请求似乎无可厚非，但是，由谷歌的算法（可能还有谷歌的律师）来判断公众对什么感兴趣，这是合适的做法吗？

早在1890年，当那两位能干的律师申诉“隐私权”的时候，他们感兴趣的其实是个人对人格的根本占有权。谁愿意涉及自己隐私的照片被他人公之于众，而且自己没有任何发言权呢？他们认为，法律应该要求人们待人以仁。法律保护“隐私权”的目的是维护民众的尊严，而不是倡导言论自由。当时的人们普遍认为不受限制的自由将导致大众变得暴虐，自由是邪恶的东西。

加州大学伯克利分校的保罗·施瓦茨（Paul Schwartz）和科隆大学的卡尔–尼可拉斯·派费尔（Karl-Nikolaus Peifer）合作发表了一篇内容深刻的论文，对隐私权概念与人格概念在法庭上能否发挥保护作用的问题进行了探讨。两位律师在文章中介绍了两本书，其中一本是美国畅销书作家写的“披露隐私”的回忆录。这本回忆录的作者回忆了她与阴痛斗争的经历，疾病给她的身体和心灵造成了双重伤害，也破坏了她与前男友的关系。她并未在书中提及男友的姓名，对他的生活细节也做了一些修改。但是，她的前男友说，他的朋友和生意伙伴都知道他们俩的关系，因此，她对性生活的描写是对他“人格的严重侮辱”，并且“败坏了他的名声”。法官也认为，这位前男友的身份在书中是可以辨识的，而且他的名声确实受到了严重损坏，但法官断言，公众从回忆录中得到的好处远大于前男友从中受到的伤害。根据公开的特性，只有一个小圈子里的人可以确定他的身份，这一点是法官最看重的理由。另外一本书是在德国出版的自传体小说，以略加掩饰的文字描写了作者的前女友和她的母亲。尽管小说有“惯用的声明，宣称书中所有人物都是虚构的”，但是一

位德国法官认为，只要是认识这位前女友或其母亲的人，都会看出她们就是小说中人物的原型。对此法官的判决是，由于这位前女友的性生活显然是隐私信息，因此她的权利确实受到了伤害；而她的母亲在小说里的活动还涉及其他人，原本就是公开的。法官禁止公开这位前女友的性生活情况以取悦公众，这本小说因此被禁。

联想到我的朋友在脸谱网上贴出屏幕截图的行为，你可能会认为这两个判决有点儿奇怪。不妨假设那张屏幕截图是真实的，那么法官在隐私权与人格权之间如何取得平衡呢？交友程序上的聊天内容可被视为隐私信息吗？如果在截图上隐去不当信息发送者的姓名与照片，结果是否会不同呢？

我之所以提到前面这两个判决，另外一个原因是公众受益与个人受害之间的权衡结果在判决中起到了重要的作用。显然，社交数据的指数增长趋势带来了前所未有的机遇。对个人造成的伤害达到何种程度，才会超过大众从中获得的好处呢？潜在约会对象的个人资料编辑记录有可能被用户不经意或恶意地分享给好友或同事，以致名声受损，你是否愿意因此放弃查看这些信息的权利呢？在包括财务管理、就业、教育、医疗在内的多个领域里，越来越多的数据服务商正在为我们的决策活动提供支持，因此，我们需要开发更先进的工具，对这些权衡结果进行评估。

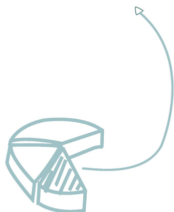
科学家兼获奖科幻小说家戴维·布林指出，似乎所有人都想自己享有隐私权，而自己的交友对象有责任心。这两者是无法兼得的。因为隐私权是一种幻象，所以我们必须让自己更有责任心。对好友负责任，就是一个好的开始。



第3章
社交图谱与信任系数

社交网络中你的身份与名望

你认识谁？他们认识谁？你又信任谁？



一个人的友谊是其价值的最佳量度之一。

——查尔斯·达尔文（Charles Darwin）

不久前，一位招聘人员在网上联系了我的一个朋友的朋友［下文把这位年轻女性称为“丽贝卡·戴维斯”（Rebecca Davis）］，询问她是否愿意接受一份新工作。这位招聘人员注意到了丽贝卡发布在领英网上的个人资料。从资料来看，丽贝卡似乎是一位年轻的专业营销人员，事业正处于上升期。同事对她交口称赞，她之前在一家很有声望的硅谷公司实习期间也得到了好评。但是，实际情况却令人尴尬。丽贝卡并不是一个真人，而是我的这位朋友虚构出来的。他的目的是验证在网上创建并维护一个虚构人物，难度到底有多大。

人们在若干社交媒体平台上都能发现丽贝卡的踪迹。她诞生于脸谱网，并且肩负着一个艰巨的任务：向现实世界中的陌生人发出加好友请求，并得到对方的同意。她可能联系过那些迫切需要更大粉丝群的三流名人，或者那些希望比同学拥有更多虚拟世界好友的年轻人。但是，如

果她不加选择地添加各种类型的好友，算法很容易就会注意到她。因此，丽贝卡精选了一些目标，然后向他们发送了加好友请求：

> 你好！我叫丽贝卡。我特别喜欢这个名字，这个名字太适合我了。我想你也有同样的感受吧。因此，我希望同脸谱网上所有名叫丽贝卡的人成为好友！

她还在脸谱网上搜索“丽贝卡思”“贝吉斯”“贝卡思”“丽巴斯”等名字，然后发出类似的加好友请求，因为我的那位朋友知道，人们肯定会把它们视为相似的名字。

令人吃惊的是，丽贝卡很快就建立了一个规模不小的社交网络，不仅有同名的“丽贝卡们”，就连他们的好友也加入进来。她开始收到来自这些好友的生日祝福，她也会像其他用户一样，只要脸谱网通知她某个好友的生活中发生了某件事，她就会做出响应。她偶尔还会更新自己的状态，上传一些食物照片，就约会与工作问题寻求一般性建议。脸谱网的算法与丽贝卡的好友都没有发现她不是真人，因为她的行为与真人没有区别。通过这些好友和信息，丽贝卡建立了数字身份。

最后，我的朋友觉得丽贝卡可以在领英网上建立个人档案了。根据她的出生日期和脸谱网贴文，她应该已经大学毕业，正在找工作。至此，她拥有了电子邮箱和脸谱网账户，足以证明她在社交网络中的存在。我的朋友先给了她一个实习生的身份和一个初级工作岗位，然后让她迅速地晋升了一次。

但是，在专业的招聘网站上虚构一个可信的职业发展轨迹，比在脸谱网上建立个人资料要难。而且，领英网为用户推荐的好友在某公司的工作时间，还会与你在那里工作的时间有重合。不过，有10多个在丽贝卡声称自己工作过的公司里就职的人将她加为好友，甚至还有几个人对她的工作给予了肯定。是他们认错了人，以为她是另外一个名叫丽贝卡的真人，还是因为他们急于扩大自己的社交网络，所以没有仔细辨认就接受了丽贝卡的加好友请求呢？无论如何，丽贝卡让人深信不疑的工作经历以及广泛的社交网络，足以吸引招聘人员的眼球。

众多互动交流信息的存在，通常足以证明某个人是真实存在的。丽贝卡的个人资料就是一个例证。拥有脸谱网或领英网账户的虚构人物比

没有这些账户的普通人更加真实可信，为什么？我们先将与脸谱网账户有关的数据分成5类：

1. 账户名和密码，这些信息可以帮助你登录脸谱网，还可以帮助你利用脸谱网登录程式码登录其他网站和应用程序。

2.个人资料里关于你的特性信息，例如你的家乡、居住地、电话号码、学校、工作单位，以及性别认同与性取向等。

3. 既有好友与好友群列表。

4. 你的贴文、评论和点赞，即你分享给好友的数据。

5. 你与好友在贴文、评论方面的交互，即你和好友在脸谱网上交互时共同创建的数据。

相对来说，前两类属于静态数据（即不变化或者不经常变化的信息），不接受其他人的评论。剩下的三类数据（我们的社交与交谈）每周、每天、每个小时甚至每分钟都在变化，它们是以对话的形式刻意创建的。

当然，有的社交与交谈会透露出更多信息。脸谱网、领英网等数据服务商特别善于测量、汇总、分析我们的社交网络和交流模式，从而为我们提供质量更高的推荐意见，包括我们想要认识或者关注的人选。几千年来，人们一直在收集信息，用于判断谁的建议值得一试，谁说的话可以信任。现在，我们的人际交往已经延伸至世界各地，这个变化正在改变我们的决策行为。对数据服务商的透明性和主动性要求越高，分享个人的社交数据给我们带来的价值也越高。

有人认为，“一个人的友谊维系时间的长短”可以反映这个人的特点。对于这个观点，达尔文深表认同。但是，在社交数据大显身手的今天，维系时间只是友谊的量度之一。如果你很少与人在网上交流，数据服务商就会发现你的这个特点。如果你同另一个人的关系还不如鼠标的点击声那么有诱惑力，这能说明什么呢？事实证明，从中可以看出你的很多信息。

大数据时代的人际关系

你也许已经发现，在我介绍的5类脸谱网数据中，前两类（即姓名和个人特性）传递的是传统意义上的个人身份信息。在某些情况下，这些数据需要由权威机构加以确认。例如，你在申请驾照时，政府机构会根据相关记录核查你的姓名、出生日期、相貌等信息。为了确认你的身份，某些数据服务商可能会要求你出示由政府发放的一系列身份证明。不过，也有一些数据服务商不断开发各种形式的身份验证方式，比如根据用户的社交关系和人际交流的结构与规律验证用户的身份。

身份不仅属于个人所有，它还具有社会属性。很多时候，我们通过行为和交互表明自己是多个群体的成员，并借助与其他人的互动，建立我们的身份。人类学家罗宾·邓巴（Robin Dunbar）提出，人类语言源于为亲朋好友“梳理毛发”的需要；闲聊对双方的抚慰作用，比帮对方捉毛发中的虱子更明显。换言之，闲谈并非毫无意义，它也有可能起到帮助作用，拉近人们之间的关系，为社交圈提供各种新闻。闲聊还有利于分享有价值的数据，反映群体成员遵守社会规范的情况。我们会将行为不端的人赶出去，提供积极的反馈信息以倡导良好的行为。在建立人际关系方面，每个人都是行家。

邓巴认为，这些“梳理”工作已经演变为由我们自己完成，而人受到认知能力的限制，最多可以与大约150个人同时交往。大约400万年前，人类的祖先与其灵长目近亲分道扬镳。从那以后，我们的大脑、身体和工具不断进化和发展。我们与黑猩猩、猿猴不是同类，尽管我们的DNA（脱氧核糖核酸）有很多相似之处。我们可以在一天之内跑到地球的另一端，也可以与几千英里之外的人视频聊天。移动技术和社交技术的发展为建立、维护人际关系提供了新的可能。通过数据服务商的努力，数百万人的个人资料呈现在我们眼前。这些人有的已经是我们的朋友，有的可能即将成为我们的朋友。马克·扎克伯格提出了“社交图谱”（social graph）这个概念，用它来表示人们在脸谱网上建立的人际关系。脸谱网利用算法分析这些人际关系，以便为用户推荐好友或信息。这个表达的来源是数学领域中研究成对关系的一个分支——图论。从本质上讲，社交图谱只有一个，你生活在其中的一个领域中，这个领域就是你的社交网络。不过，由于脸谱网的用户超过10亿，整个社交图谱已经接近数字化状态了。这是一个令人惊讶的变化。在现代通信手段问世之前，我们能够研究的最大型的社交网络往往是村庄、学校或者公司这样的规模。

对比互联网诞生之前的社交网络，你会发现，在脸谱网等社交平台上与好友交流时可用的数据实在是太多了。20世纪30年代，精神病专家J·L·莫雷诺（J. L. Moreno）开始为人际关系和人际影响创建“社会经济”图谱。他的一个案例——纽约州问题女生“出走成风”原因调查，引起了人们的关注。决定逃学的女生住在不同的宿舍，有不同的家庭背景。学校负责人为此头疼不已，向莫雷诺求助。莫雷诺把这些问题女生之间的友谊绘制成图，并且标示出她们之间的友情以及每个人的活跃程度与智力水平。有的女生是焦点人物，可以把其仰慕者吸引进入她的社交圈。有的好朋友组合同样看重她们之间的友谊，一条共同的纽带将她们拴在一起。莫雷诺认为，出走的女生不仅有共同的好友，也有相同的人生态度和价值观。

莫雷诺的分析表明，社交图谱会影响人们的决策行为。但是，在风险更高时，情况仍然如此吗？社会学家道格·麦克亚当（Doug McAdam）的一项研究给出了一个答案。1964年，一些社会活动积极分子申请参加著名的“自由之夏”计划。事后来看，参加这次活动的人“受到了肉体与情感的双重创伤”。三名积极分子在到达密西西比州的短短几天内，就遭到绑架、谋杀。晚间新闻明确表示，这次活动非常危险。于是，一些申请参加活动的人退却了：在这个夏天到来之前，申请并获准参加活动的人中，有25%最后没有参加行动。为了弄清这些人提出申请的动机，麦克亚当认真地研究了他们的申请书。他发现，最终前往密西西比州的那些人大多与另外一名“自由之夏”参与者或民权活动积极分子有非常密切的联系，而且这个动因在重要程度上超过了他们之前参加任何类似活动的经历。

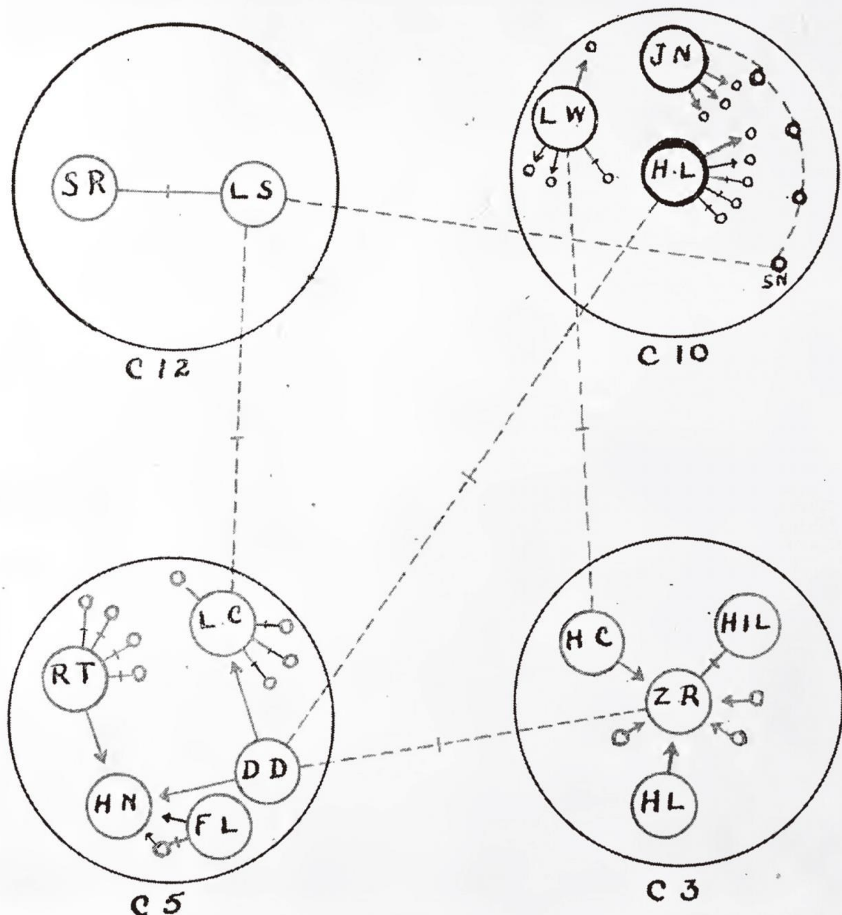


图3–1 雅各布·莫雷诺早期完成的揭示出走女生与未出走女生之间关系的问题女生社交图谱
资料来源：选自J·L·莫雷诺的《谁能活下来？——人际关系问题新解》。本图片经乔纳森·D·莫雷诺授权使用。

更广泛地说，人际关系和交流规律图谱可以为信息传递和聚拢专门人才提供帮助。商学院老师与企业管理顾问习惯于将正式的组织结构图（谁向谁报告）与非正式的信息沟通流程（谁该去找谁，解决哪些问题）放在一起，比较孰优孰劣。通常，社交网络分析可以发现沟通过程中的瓶颈，并为改善公司管理开出良方。例如，IBM（国际商用机器公司）知识基础组织学院研究发现，某个石油公司的一位中层管理人员"因专业技术背景深厚、反应迅速而名声在外"，因此"他收到的项目邀请以及他参与的项目数量迅速激增"。最终，这位管理人员不堪重负，公司的多个项目也无法如期完成。

这些早期的社交网络分析常常依靠访谈、调查，以及偶尔对实验对象的直接观察来获取数据，但是在现代交流技术问世之后，研究人员便

开始利用社交图谱的数字痕迹。甚至在手机出现之前，分析电话记录就已经是研究社交网络的最简单方法之一。电话公司帮你接通电话之前，必须知道你想要呼叫的号码；收费时，还必须知道通话时长。因此，电话公司都精于数据追踪之道。1991年，MCI通信公司宣布实施“亲朋好友”计划，目的是从占据2/3长途电话市场的美国电话电报公司（AT&T）那里抢夺客户。MCI通信公司请客户列出20个电话号码，许诺他们拨打这些长途电话时享受话费折扣。如果他们每个月的长途电话费超过10美元，就可以享受8折优惠。他们列出的人选可以不是MCI通信公司的客户，MCI通信公司会负责联系这些人，向他们兜售这项计划。在两年之内，有1 000万客户加入了这项计划。通过经济刺激，MCI通信公司诱使客户说服他们经常联系的人投入了它的怀抱。

电子邮件为社交网络分析提供了另外一个数据来源。20世纪80年代末，科罗拉多大学博尔德分校计算机科学教授迈克尔·施瓦茨（Michael Schwartz）希望解决如何在互联网上找到志趣相投的人这个难题。［几个月之后，蒂姆·伯纳斯·李（Tim Berners-Lee）提出了万维网的架构，大幅降低了互联网搜索的难度。］迈克尔分析了15所大学和研究实验室（包括加州大学伯克利分校和太阳微系统公司）在两个月时间里发送和接收的电子邮件。他利用电子邮件的“标题”数据（仅包含发件人和收件人），绘制出包含50 834名研究人员的社交图谱，展示未来可能开展的合作。当时，分析100万条信息需要收集为期两个月的电子邮件数据。今天，脸谱网信息与通信程序WhatsApp每秒钟就可以处理10万条信息。

这些例子说明，社交图谱就是人与人交往搭建而成的网络。用计算机科学的术语来说，社交图谱是由节点（每个节点代表一个人）组成的，节点之间通过链路或者棱连接。有的人喜欢同在某些方面与自己类似的人建立联系，包括在状态（特性）或价值观（态度）上具有相似性。这种现象叫作同质性（源于希腊语，意思是“爱好相同”）。人们之间的交互形成链路结构，随着交互的增多，两个人之间的棱的权重就会增加。有的人经常联系的人不多，他的棱的数量较少，颜色深重；有的人与很多人都保持浅显的联系，他的棱都比较细，呈发散状。

社交网络的结构可以透露出大量信息。有的社交网络节点比较少，彼此之间大多是相连的，紧密地凝聚在一起；有的社交网络节点较多，彼此相连的较少，呈现出稀疏、不连贯的形状；还有的社交网络包含若

干节点群，群内节点大多相连，群与群之间的联系则比较少。紧密凝聚的社交网络表示信任程度较高，具体原因在于该网络中可以直接了解其他节点个人信息的机会较多。稀疏、不连贯的社交网络表示信任程度较低，因为信息流通的机会不多。松散、多样化的网络一般表示直接了解群外节点的机会较少，但是为人们通过与外围的联系发现新想法和机遇提供了便利。有时，这些情况会随着时间发生变化，例如你与父母之间的关系会随着你在不同人生阶段的身份变化而变化。

同真实生活中一样，互联网上各种关系的平衡性也不尽相同。造成这种状况的原因有时是社交网络的规则与习惯。在推特上，你可以关注或提及你感兴趣的任何人，无论这个人对你是否感兴趣；终结同其他用户之间不愉快交互的终极手段是屏蔽这些用户。联系是单向的，对话常常也如此。但是在脸谱网上，联系是双向的，需要双方共同确认。

共同确认的友谊并不表示平衡性一定很好。两个人之间交互的次数与方向还可以表现出关系的其他特点。随着两个人不断交互，两个人之间的棱就会逐渐加粗，但是，这条棱在一个方向上的权重可能大于另一个方向上的权重。例如，你喜欢定期联系好友马克，因为你有一个创建科技初创公司的想法，希望马克为你引荐潜在投资人。你们偶尔碰面，偶尔还会互通邮件，但是几乎没有深入交流的机会。因为迫切希望吸引马克的注意，你对他的状态更新发表评论，为他的照片点“赞”，还直接给他发信息。你每联系他10次，马克就会回应一次。这似乎可以区分你们的兴趣程度。不过，最终联系上马克之后，他对你直接发给他的信息做出了回应，而你对他的贴文做出的评论，以及你为他的度假照片点的“赞”，却如同石沉大海。马克的回应反映出他的兴趣，以及兴趣的程度。脸谱网的算法正是根据交流中的这些细微之处，确定应该提示你注意哪些贴文。

斯坦福大学的社会学家马克·格兰诺维特（Mark Granovetter）研究过人际关系中联系纽带的强度。在他于1973年发表的开创性论文“弱关系的强度”中，他对关系的强度进行了定义，即时间长短、情感强度、亲密性（相互信任的程度），以及相互帮助的程度等的综合体。网络中人们交换的不仅是情感和信息，还有影响力和帮助。

过去，社会学家深入村庄和公司，实地调查人们之间的交流情况。营销人员希望通过打折活动，引诱顾客推荐潜在客户。现在，脸谱网等社交数据公司正在实时利用我们的数字痕迹，在为我们提供服务的同时

也在改变人际关系的性质。

“动态信息”功能与“分享所爱”计划

在创建了数字身份之后，我们开始重新定义友谊。过去，我们必须把大部分时间花在采集食物、交朋友方面。现在，食物已经不成问题，与朋友联系也更加方便，但我们在如何安排时间方面却没有取得任何进步。

要保持友谊，就要不断地披露信息。发展心理学家发现，在5～9岁这个阶段，愿意共享玩具的孩子都视对方为朋友。经过缓慢的演变，友谊被定义为双向合作，代表一种平衡的互惠关系：孩子愿意和伙伴分享自己的玩具，前提是伙伴也愿意跟他分享玩具。随着年龄的增长，我们可能会继续分享“玩具”，但分享更多的是秘密。两个人你来我往，通过一次次交谈了解对方。一个人分享一点儿通常秘而不宣的信息，然后另一个人就会接过接力棒。“相互间持续不断，且进一步展示自己的个性”是建立友谊的一种普遍做法。据报道，某心理学研究团队可以诱使两名陌生人产生亲近的感觉，方法是让他们回答一系列问题并彼此分享答案，这些问题会逐渐深入地涉及隐私内容。

“秘密”这个概念本身就暗含披露该信息可能会带来危险的意思。社会学奠基人之一格奥尔格·齐美尔（Georg Simmel）指出：“人与人之间各种关系的特征，就是在这种关系中相互保守的秘密有多少。”在建立友谊的过程中，我们必须加深对另一方的信任，直至克服这种危险。如果一个人选择对另一个人保密，尤其是这个秘密在重要程度上与另一方披露的秘密大致相似，他们之间的关系就会失去平衡，交换信任的过程可能会中断，双方之间的链接纽带也会减弱。

这种不平衡性可以通过社交图谱中的棱表现出来，也会影响我们与数据服务商之间的交互。如果一方分享信息之后，对方同样分享一些信息作为回报，前者就更有可能再接再厉，继续分享自己的信息。我们知道，最成功的数据服务商鼓励用户贡献原始数据，如果这些数据可以明显改善它们的数字服务与产品，它们还会给予用户奖励。对等性可以有效地协调数据提供者与数据服务商之间的利益关系，原因在于将欲取之，必先予之，这是人类的天性。

当交流技术局限于发送烟火信号、街头广场闲聊或者信件往来时，当传感技术仅限于人的眼、耳、鼻、嘴和皮肤时，人的影响力只存在于局部范围内。这种情况在20世纪发生了变化。现在，广告与营销通过无线电、电视与互联网，在全球范围内影响人们的生活。过去，我们听从亲友和隔壁邻居的推荐，有时某位演员会扮成“有辨别能力的顾客”，和我们分享他在高端连锁酒店的“五星级体验”。今天，脸谱网根据10亿人创建的数据，为我们提供推荐意见。我们在社交图谱上的邻居发挥了重要作用，帮助我们筛选数据，为我们提供个性化建议。

大体上讲，脸谱网是我们与朋友交流的一个平台。你的利益与它是一致的。脸谱网希望把你感兴趣的东西推荐给你，这样你才会再次登录。当你再次登录时，脸谱网会通过数据挖掘程序进一步了解你（你会花时间看哪些贴文，推荐哪些好友时你会点击查看）。当然，脸谱网的算法还可以更精准地向你投放你感兴趣的广告，帮助网站赚钱。

通过分析你点击鼠标、浏览网页等行为，脸谱网可以了解你是如何分配注意力的，记录你对哪些人、哪些东西感兴趣，以及你的兴趣如何随时间发生变化。你会花一个小时的时间浏览一位新添加好友贴出来的假期相册，还是更愿意花一分钟的时间为姐姐家的小宝贝近期拍摄的照片点赞呢？这些数据是一个信号，体现了在特定时间里你赋予人际关系的优先等级和兴趣强度，比你事后回想你是如何分配时间的这个方法精准得多。而且，在为输入数据分类时，这些数据也远比关于社会责任的各种规章制度有用。脸谱网在为你定制“动态信息”（News Feed）时，主要依据的就是这些真实的信号，也就是你的注意力和兴趣留下的踪迹。

“动态信息”功能根据用户的人际关系网发布社交信息，以引起人们的关注。这项服务建立了一个积极的反馈回路：内容与用户非常匹配，用户知道会有更多的人为该内容点赞，因此发帖者，或许还包括看到该贴文的其他人，会受到鼓励，愿意分享更多类似的信息。与之相反，否定响应通常仅限用户本人知道。脸谱网让用户做出选择，可以通过两次点击，表明“以后少给我发一些这样的贴文”，但是脸谱网不会通知发帖者有人不喜欢这样的内容。“动态信息”还允许用户处于“持续性部分注意”的状态。这个说法是琳达·斯通（Linda Stone）这位有远见的科技领袖创造出来的，用来形容我们总是在观察好友，同时也被好友观察的那种情况。琳达认为，持续性部分注意“需要人为地制造持续的危机感”，

让大脑始终保持清醒，扫描从“动态信息”以及其他数据源不断涌来的信息。不过，脸谱网还会从“动态信息”中过滤掉大量信息。想象一下，如果你真的可以看到好友发在脸谱网上的所有信息，你的生活将会变成什么样？

脸谱网对这些交互活动的分析非常精细。据脸谱网前技术主管丁周（音）介绍，除了根据所发信息的类型（是状态更新还是上传照片）决定两人之间关系的权重以外，脸谱网还开发了基于本体的话题检测与跟踪技术，对各种交互行为进行分类。例如，在现实世界中，我请医生为我提供医疗建议，但我不会就如何修电脑咨询他的意见。在脸谱网上，我们也会先考虑贴文作者能否激发我们的兴趣，是不是专家或话题相关领域的权威人士，在此基础上再对他们的贴文做出回应。

不过，脸谱网只是我们可以利用的数字交流平台之一。手机和讯佳普上有拨打和接听的电话记录，告诉我们谁呼叫了谁，通话持续了多长时间。事实上，如果你使用讯佳普打工作电话，讯佳普就可以通过分析电话数据，找出你在工作上有哪些联系人，包括你最关注哪些客户或同事。

你无须怀疑，一些公司正在利用这些信息，探查你与哪些人有联系。在MCI通信公司的“亲朋好友”计划大获成功之后，其他公司也在努力收集可以用来吸引、影响客户的社交图谱数据。2001年，亚马逊（先于脸谱网）启动了“分享所爱”计划。在你买了某件产品之后，亚马逊会询问你，在你认识的人当中是否有人希望了解你的这次购买行为。如果你的回答是肯定的，而且分享了这些人的邮箱地址，你的这些好友就会收到一封邮件，告诉他们如果购买这件产品可以享受9折优惠。这绝不只是一个无私奉献的机会：除了可以炫耀你的良好品位与新锐风格以外，如果收到邮件的任何人在一周以内购买了这件商品，你就会得到原价10%的退款。由于推荐双方都可以享受折扣，因此有很多人都愿意向亚马逊提供真实的邮箱地址。提供虚假或失效的邮箱地址，你不会得到任何好处，从而避免了早期电子邮件推销计划的一个弊病。结果，这个计划的销售成交率远高于其他促销手段。“分享所爱”计划为亚马逊提供了一个工具，帮助他们根据回报与社会认同等心理学原理，为客户建立经过核实的社交图谱。

与之类似，AT&T在将一项新服务推向市场时，决定做一个A/B测试，以了解根据传统的细分市场和社交图谱做宣传，哪一种方式可以吸

引更多的顾客。AT&T的营销团队对顾客十分了解，知道他们住在哪里、与AT&T的合作关系保持了多长时间、选择了哪些营销计划、是不是AT&T的忠实客户、是否会在多个电话公司之间摇摆不定。为了比较社交图谱与其他营销手段孰强孰弱，AT&T同意做一个非常简单的尝试：如果你经常通过电话联系的某个人已经是该公司的客户，AT&T就会向你推销这款新服务。同“分享所爱”计划不同的是，你不知道自己为什么会收到推销信息，因为这些信息中没有提到推荐人。不过，如果你的“电话图谱”中已经有人先行一步使用了这项服务，那么你也同意购买的可能性会增加大约4倍。

同质性有助于彼此相似的人建立起更紧密的关系，那么你是否会因为同质性而购买这项服务呢？你购买这项服务，是不是因为营销资料使你想起你的朋友对这项服务的肯定呢？或者，你更关注朋友的看法，是因为你之前在营销活动中听过类似的评价呢？数据无法回答这些问题。但是，令人吃惊的是社交图谱的宣传效果远强于细分市场营销法。预测你的兴趣，不是为了知道你是谁，而是为了知道你认识谁。

为拥有数据的人提供服务

你认识谁，这在职场具有非常重要的意义。马克·格兰诺维特发现，人们在找工作时，宁愿通过熟人介绍，也不愿意借助好朋友、职业介绍所或招聘广告。这个发现或许并不奇怪，因为熟人的数量远多于亲密好友的数量。但是，不仅如此，他还发现通过社交图谱上的弱关系找到的工作更令人满意，薪水也更高。

格兰诺维特的这些发现要追溯至20世纪70年代。那时没有电子邮件，也没有招聘网站，信息传播不仅速度慢，而且成本高。很多专业人士都是通过“罗拉代克斯名片盒”找到工作的。利用名片盒里精心收藏的联系方式，他们与经纪人签约，得到新的工作机会。你也许仍然认为你的罗拉代克斯名片盒是你最珍贵的职业资源，你也许仍然认为你所在的公司应该更努力地按照员工们多年来积累的客户名录维护客户资源。大多数管理人员一直坚信，这种信息优势是企业成功与否的分水岭。

领英网的一个宗旨是为你与你的业务关系（包括强关系和弱关系）创造便利的交流条件。格兰诺维特感兴趣的是那些帮助人们取得现有成

就的弱关系，而领英网则致力于帮助用户找到那些有助于他们实现目标的弱关系。假设你的目标是搞定一个新客户或者找到一份新工作，而且你知道相关负责人的姓名，但是你们没有私交，你也不知道你的联系人当中谁与他有私交。如果你搜索他的姓名，领英网就会告诉你，你需要经过多少环节才能找到引荐人。如果你选择使用付费服务，领英网就会告诉你哪些人能帮助你以及他们的姓名。这项服务旨在为需要数据的人提供便利，但它会导致一种严重失衡的关系，因为那些需要数据的人可以得到帮助，却无以为报。事实上，在领英网刚开始提供这项服务时，用户们抱怨他们收到的联系请求太多了，以至于有时候他们认为这些信息与垃圾信息无异。一位时事评论员称领英网“正在利用人脉广泛的人所拥有的高超的社交技能，为那些人脉少的人填补社交短板”。

领英网战略计划部前副部长埃伦·列维（Ellen Levy）说，我们在设计网站时，直到把服务对象确定为那些拥有数据的人，而不是需要数据的人，才取得了真正的突破性进展。埃伦拥有斯坦福大学的认知心理学博士学位，从她后来加入领英网以及她本人的履历可以看出，她在上学期间对如何在时间不足的情况下获取有效信息这个问题有过研究。埃伦说：“最不适合建立人际关系的时间是在你有所求的时候，需求会把交友变成交易，这两者根本不是一回事儿。建立人际关系的最好办法是帮助别人，而且没有任何不可告人的动机。”如果领英网只是罗拉代克斯名片盒的简单复制品，人们就没有任何理由使用它。如果登录领英网后能体验到的唯一特有服务就是有一群陌生人向你发出联系请求，那你更会避之不及。要从用户那里获取数据，领英就必须先给用户提供数据。

在为用户提供服务的过程中，领英网面临的难题是，如何鼓励用户创建并分享更多的职业社交网络数据。通常，人们只有在需要帮助时才会联系职场中的熟人，这与埃伦的建议正相反。必须为用户提供理由，他们才会增加与这些熟人联系的频率。于是，在这些熟人学习了新技能、掌握了新经验、换了新工作，或者庆祝工作周年纪念日时，领英网就会向用户发送提示信息。网站还举办论坛，让用户发帖子，展示自己的专业知识，或者对相关新闻进行评论。这些活动为用户之间进行交流提供了更多的理由。如果用户允许领英网接入他们的日程表，领英网就会将约见对象的相关信息提前发送给用户。此外，领英网还提供职业技能培养服务，根据个人资料的“完整度”和质量，以及联系人、点赞、发帖数量与其他网站活动，对用户进行相应的奖励性培训。

领英网给用户的另一个选择权是让他们有机会为联系人的职业技能点赞。点赞表示你愿意支持他，即使这对你来说没有什么明显的好处。在某些情况下，点赞甚至可能会激发对方以同样的方式回报你的想法。在领英网的社交系统中，不同的交互有不同的权重。大多数人都认为，一封精心撰写的推荐信的效果好于为某一项技能点赞。此外，为你点赞的人是否受欢迎、是否有名气，以及你们在公司的工作时间是否有重合、你们在相同部门还是不同部门，这些因素都非常重要，而且都可以赋予一个权重。这些特性的重要性还会随着时间逐渐降低，如果你与某人结束共事关系已有10年，这个人的权重就会小于你现在的同事。

如果我们静下心来思考，就会发现这是领英网创造收益的一大法宝。提供网站数据，帮助企业招聘人员发现并招揽潜在员工，这项服务为领英网创造了60%的收益。此外，领英网还向企业客户提供潜在客户管理工具、经济与就业趋势分析等服务。领英网利用的原始数据都来自个人，而不是企业的人力资源部门，因为企业不愿意把高级人才的经验或者素养透露给竞争对手的人力资源部门。正是出于这个原因，领英网向愿意分享工作和职业数据的个人提供免费服务，包括向他推荐他可能认识或希望认识的人，以专栏和幻灯片演示的形式提供业务咨询建议，告诉他谁在浏览他的个人资料等。

如果领英网提示你有人浏览了你的个人资料，你可能很想知道这个人到底是谁。但是，如果你是一名管理人员，想在面试某人之前进一步了解他的经验和兴趣，或者想进一步了解竞争对手的管理层，为“挖墙脚”做准备，那么你肯定不希望他们知道你浏览过他们的个人资料。在浏览他人的个人资料时，用户可以在显示自己姓名与不显示姓名这两种状态中轻松切换。你可以选择匿名浏览他人的资料，而只显示你所在的城市或行业。但是，如果你选择显示全名，你就可以看到哪些人正在查看你的资料。根据他们选择的身份显示方式，你可以看到他们的姓名、城市或行业。如果你是匿名状态，就无法看到上述数据。当然，无论你选择哪种设置，领英网都会记录你浏览过的所有资料。但是，你的选择将决定你的哪些信息可以显示出来，而这些信息的精细程度又决定了你在浏览他人的个人资料时可以看到哪些信息。

在和交友网站Skout合作期间，我考虑过什么程度的透明性对用户最具吸引力。用户在Skout网站上建立个人档案是不收费的，这是因为网站的用户越多，就越容易吸引有网上交友欲望的人。我们希望尽可能

地扩充用户群。我们还发现，如果用户需要付费才能浏览他人的个人资料或联系他人，他们就会尽可能少地使用网站服务。不过，我们还是有利可图的，因为用户用鼠标点击他们感兴趣的其他用户的行为，向我们发出了真实的信号。用户可能甘愿掏腰包，了解有哪些人对他们产生了兴趣但还没有通过发送信息的方式做出明确表示。最后，我们推出了一项高级功能，允许付费用户查看谁点击了他的照片以及查看他的个人资料的认真程度，例如，他看了多少张照片，是否或者何时又浏览了一次。我们还研究了是否要提供VIP（贵宾）用户服务，让愿意多支付一个月费用的用户可以用秘密模式查看他人的个人资料，也就是说，让VIP用户享有隐藏真实信号的权利。

我在脸谱网上贴照片，是因为我估计好友可能会看到这张照片。如今，我们唯一能做的就是根据好友的点赞与评论情况推测他们是否对这张照片感兴趣。我希望自己有权决定哪些人可以看到这张照片，就像在“访客留言簿”上签到一样，他们在这张照片下方签名，表示他们愿意让我知道他们看过，然后才可以看到照片。脸谱网是否应该把这个决定权交给我呢？如果我邀请朋友来我家做客，咖啡桌上放着一本相册，朋友就会知道，无论她是瞟一眼、快速翻看几页、认真地研究一番，还是掏出手机拍一张我的照片留存（这种假设令人难以置信），我都会一清二楚。脸谱网让我们看到朋友的点赞和评论，却不让我们知道谁看了我们的照片，尽管它有这方面的数据。脸谱网上的好友可以下载我上传的任何照片，而我却一无所知。我希望领英网资料浏览方面的对等性能被更多的数据服务商采纳，同时这种对等性可以应用到更多类型的内容上。

怎样才能看到你与好友的关系伴随着好友留下的一个个数字“脚印”发生变化的过程呢？如果你知道好友看过你的所有照片，但既没有评论也没有点赞，那么你看他照片的可能性会增加还是减少呢？好友花多长时间看你的照片是否重要？如果知道有人在观察自己，大多数人都会约束自己的行为，减少点击和浏览照片的数量。包括脸谱网在内的网站和应用程序都希望用户尽可能多地交互，这样它们才能得到更多的反映用户真实兴趣的数据。鼠标点击和网页浏览是反映用户兴趣和关注点的真实信号，也有助于数据服务商为用户推荐更合适的内容，包括新闻和广告。

数据服务商在思考提供哪些服务以及如何将这些服务推荐给用户等

问题时，都会考虑到社交数据离不开某个生态系统这个问题。生态学中的生态系统是指在某个环境中相互作用的一群有机体。很多生态学研究者认为，要处理好生物之间的相互关系，不能仅从个体层面考虑问题，还必须考虑整个生态系统的健康状况。如果试图改变某一个个体或者种群的状况，就有可能导致整个生态系统陷入紊乱状态。想一想托马斯·奥斯汀（Thomas Austin）的教训吧。为了周末打猎时能有更多的收获，托马斯在他位于澳大利亚吉朗市的庄园里养了24只英国兔子。现在，兔子成为当地的一大祸害，不仅导致大范围的土壤侵蚀，还导致本地物种遭到破坏。

利用社交数据帮助人们择偶时，就需要维护好整个生态系统的健康状况。每个人都是独一无二的，但每个人每天集中注意力的时间都无法超过24个小时。所以，你喜欢或者认识的人不一定会对你做出回应。亚马逊从事的业务是推荐和销售批量生产的产品，但它无法为某个人制造出多个副本。如果脸谱网提示你或许应该把艾米加到你的好友名单中，那么你注定会失望，因为艾米的好友数量已经达到了上限。如果某个交友应用程序告诉你，你或许会对约翰感兴趣，那么你也逃不了铩羽而归的下场，因为整个一周约翰都排满了约会，而且他对这些约会对象的兴趣超过对你的兴趣。交友应用程序推荐的人选至少应该部分地回应你的邀约。在激起失望情绪这个方面，任何东西都比兔子强。如果感到失望，人们通常就不愿意创建、分享数据。既然无法获取价值，为什么还要分享呢？

我认为，动态系统研究的一些概念有助于我们更好地理解生态系统的演化过程。在20世纪60年代，物理学家发现动态系统有时会表现出一种所谓的“混沌”特性，也就是说，无论我们对某个系统初始状态的了解有多么深刻，都无法预测该系统长期运转的具体状况。混沌理论研究人员还证明，经过时间的发酵，看似微小的差别可能导致显著的变化，并产生非常巨大的影响。在某些情况下，系统可以放大随机噪声。在麻省理工学院数学家爱德华·罗伦兹（Edward Lorenz）做了题为“巴西丛林里一只蝴蝶扇动着翅膀是否会导致得克萨斯州刮起龙卷风？”的报告之后，“蝴蝶效应”因此得名。与之相似，数据服务商在设计和参数上做出的微小改变（例如，脸谱网决定不设置点“踩”按钮，不让用户看到评论与贴文的编辑记录），可能导致用户未来的行为发生明显变化，并且会影响到社交图谱的结构。

几年之后，我们可以观察一个自然实验的结果：通过比较脸谱网与微信这两个世界上最大的信息平台，了解设计上的微小区别对社交数据动态系统演化进程的巨大影响。4年时间里，微信的用户超过了5亿（其中大多数是中国人），这与脸谱网向非学生人群开放注册4年之后公布的用户人数大致相仿。由此可见，这两个平台的发展速度差不多。

不过，在用户对交流平台的需求方面，微信与脸谱网产品经理的推测却大相径庭。脸谱网源于哈佛大学宿舍楼编纂年刊作为传世档案的传统，微信的母公司腾讯最初则是一个网络游戏公司。游戏结束之后，用户就会离开。游戏的最终得分以及高分玩家的记录可能被保存下来，但并不是玩家的所有举动都有记录。这种做法后来被传承到腾讯的信息平台上。微信在设计上关注的是短暂交流，信息一旦被用户阅读，就会从公司的服务器上被删除，只有用户的设备上还留有记录。如果手机丢了，那么所有的交流记录也丢了。

这两个平台在设计上的另一个显著区别在于用户之间的联系方式。如果你收到某个人的加好友请求，但是从对方的姓名和资料照片上你无法判断这个人是谁，那么你是否可以查看对方的好友名单呢？至少应该可以查看你和对方有哪些共同好友吧？这个问题的答案在很大程度上取决于你是在哪里长大的。生活在美国的脸谱网用户肯定会想，我当然希望查看他的好友名单，因为这有助于我决定是否接受他的加好友请求。查看共同好友名单，通常可以明确他与你是前校友、前同事，还是可以将他“接受”为好友的其他关系。

微信从来不会把某个用户的好友名单透露给其他用户，这是一种看不见的社交图谱。生活在中国的微信用户肯定会想，我当然不希望他能看到我的好友名单，因为这有可能暴露我不想让他知道的信息。不经介绍，用户不可以查看他的好友的好友名单，更不用说联系他们了。

微信有一些独创性的办法，可以帮助用户在看不到其他用户的好友名单的情况下找人。例如，遇见某个人时，你可以扫描微信在她手机上生成的快速响应矩阵码（QR码，即二维码），就可以将他加为好友。用户还可以创建临时聊天群，在一群人（无论是有私交的朋友还是工作上的同事）约定见面的时间和地点时，就会用到这个功能。必须经过群成员的邀请才可以加入该群，因此群可以起到引荐人的作用，同时还是增加联系人的一个捷径。进入群后就可以看到群里的其他成员，如果想和其中某些人建立联系，就可以向他们发出加好友请求。

因为微信不让用户看到其他人的联系人名单，因此它可以利用社交图谱来确认身份。如果你忘记密码无法登录，微信就会显示一个安全码和若干用户的姓名、照片。你必须通过某种方式联系这些好友，请他们将安全码发送给你。只要有两人完成上述步骤，微信就会解锁你的账户。这种“挑战应答”式身份验证要求你证明你对自己社交网络的了解程度，安全程度远高于要求用户回答一些常规问题的做法。你母亲娘家的姓氏、你的第一份工作、你的宠物名字等常规问题，有时可以通过搜索你的贴文和你在互联网上的痕迹找到答案。为了解开你的账户，你把自己的社交网络的相关信息透露给微信，因为你希望所选好友可以迅速证明你的身份。

不过，我怀疑微信之所以不向其他用户透露联系人名单，更主要的原因可能与中国商业与中国社会赋予联系人的价值有关。在中国，微信被广泛地应用于职场人士之间的交流，在某些场合，公开自己的社交网络可能会泄露某些秘密。竞争对手查看你最近添加的联系人，就可以据此推断你的经营战略，并制定应对措施。同时，如果其他人看不到你的联系人名单，你就无须担心他们会根据“你所交往的人”来评判你。

各条棱表示的联系强度会随着时间的流逝发生变化。随着社交网络平台的不断演化，社交数据量不断增加，我们需要平台提供更多的功能，帮助我们在确保生态系统健康状况良好的前提下管理社交关系。你也许非常清楚自己经常联系的人是你的母亲还是你最好的朋友，也可能知道你对某个商家的产品不感兴趣，因此你不看他发来的电子邮件。但是，你或许不是很清楚当你在工作上需要建议时，你通常会联系哪位同事，甚至你可能都没有注意到自己已经不再查看某些好友的脸谱网更新信息了。

社交网络的相关数据经过挖掘后同样可以帮助你维护既有的关系。Skydeck网站是互联网早期创立的一家基于云技术的电话服务商，可以提供呼叫者的身份识别和呼叫屏蔽服务。该公司推出过一个试用产品，提醒用户注意他们在打电话方面的某些不好的习惯。我就曾经收到过这样的提示，说我给某个好友打电话的频率不如以前那么高了。于是，我赶紧联系这位好友，以维系我们之间的友谊。

数据服务商不仅可以提示单个用户注意他们行为的变化，还可以把他们分析整个用户群得到的发现分享给用户。例如，脸谱网发现在两个人公开恋情之前的100天里，他们在网上交流的热度会稳步上升，但

是，一旦他们修改了个人资料中的情感状态，他们在脸谱网上的交流就会急剧降温。与此同时，他们交流的内容也会发生变化，贴文和信息中出现的积极性词语有所增加。研究人员还发现了一个特殊的“签名”：根据两个人的共同好友在他们社交网络上的分布情况，就可以推断两个人是否在谈恋爱。即使他们没有明确表示“建立了爱情关系”，脸谱网也能知道真相。借助更广泛的数据源（诸如，在同一张照片中被人标签，出现在同一事件之中），数据服务商就可以推断出社交关系的强度和动态变化。

具有明显规律性的人际交流不仅出现在爱情关系之中。想象一个求职面试的情境。一位似乎胜券在握的求职者强调她曾经在某个大型公司度过了一段美好的时光，面试官可能很熟悉这位求职者提及的那位联系人。他可以给那位联系人打电话，请他评价他的这位前同事。他也可以请这位求职者说明某个数据服务商对她的职场人脉和交流模式的“鉴定”结果。这位面试官或许想知道这位求职者所说的自身优势是真是假，又或许他对这位求职者在面试中展现的优势不太感兴趣，而更想了解她在行业内的人脉情况。尽管这位求职者大谈特谈某一位联系人，但核查她在人际交流中表现出来的规律性是否与所谓的“超级联系人”相吻合，或许会对这位面试官的决策有帮助。如果吻合，就表明这位求职者适合这个职位，因为她不仅认识很多人，而且她对需要不断地与不同的人进行交流的工作甘之如饴。数据服务商的推荐意见在很大程度上取决于探索与利用之间的平衡性。这位面试官需要做出决定，在选择新聘人员时，想要实现的重要目标是增加公司现有业务联系的深度，还是努力建立新的联系？

如果需要将你的职场交流模式分析分享给你的潜在雇主，你会不会介意？作为交换，你是否想看到这位招聘经理的类似资料呢？在领英网上就有这样的交换条件：如果你想知道谁在查看你的资料，你就得同意让其他用户知道你在查看他们的资料。你是否想了解整个团队的交流模式分析呢？这些分析可以帮助你进行面试准备，让你知道如何展示自己才会让面试官认为你的加入会增强团队实力，以及你在面试过程中应该谈论哪些内容。这些方法在决策过程中可以发挥重要作用，无论你是决策人还是决策对象。

社交数据的影响力有多大

我认识的一个人（姑且叫他“乔”吧）决定尝试使用脸谱网，看看它为什么能引起这么大的轰动。乔已经60多岁了，他坚定地认为个人隐私不可侵犯。他不愿意将个人生活的相关信息暴露在互联网上，因此他注册时用了假名。他不希望现实世界中的朋友认出他，因此他没有把他们加为好友。同丽贝卡不同的是，乔没有在虚拟世界中交朋友。他在社交图谱上的节点是孤立的。毫不奇怪，乔每天上午登录脸谱网，都没有发现任何有意思的东西。各种新闻和信息都与他无关，也无法引起他的兴趣。乔在脸谱网上没有得到任何令他感到惊喜的体验，然而，他怎么可能感到惊喜呢？脸谱网与《纽约时报》是不一样的。报纸可以根据编辑的喜好，向所有人传递相同的新闻，而不用考虑读者的身份。脸谱网不是“即用型”灵丹妙药，想要从它的“动态信息”中获取数据，必须先为它提供数据。但是，乔并不清楚这些。

无独有偶，伊利诺伊大学的凯利·卡拉哈丽奥斯（Karrie Karahalios）发现，在一项关于脸谱网“动态信息”的研究中，有多达62.5%的实验对象甚至不知道脸谱网推送给他们的信息都经过了算法的处理。在实验过程中，卡拉哈丽奥斯让用户比较他们的脸谱网好友在某一天里发布的全部贴文与出现在他们“动态信息”里的贴文。部分实验对象吃惊地发现，他们的亲朋好友发布了贴文，但算法却对他们隐而不报。他们一直以为，这些联系人在脸谱网上不活跃。

为了提高数据挖掘过程的透明性，卡拉哈丽奥斯和伊利诺伊大学、密歇根大学的同僚们开发了一个名叫“FeedVis”的审核工具，帮助用户了解点赞、评论与贴文在被显示时发生了哪些变化，使他们有机会体验不同的新闻推送方式。在第一阶段，FeedVis审核工具通过比较好友网络中所有人分享的内容，即按时间顺序进行“无遗漏式”的新闻推送，向用户展示个性化的脸谱网“动态信息”。在第二阶段，FeedVis审核工具按照每名好友分享的新闻被用户的个性化“动态信息”收录的百分比，将好友分为“几乎不看”（小于10%）、“偶尔看”（45%~55%）和“大多数时候都会看”（90%以上）三个组，并向用户展示好友的分组情况。最后，用户查看好友名单之后，可以将某个内容由隐藏区移至显示区，或者在三个组别之间移动好友。此时，由用户亲自管理的最终版“动态信息”就会出现在用户面前。

不少参加过这项研究的实验对象都认为，脸谱网根据交互活动，比较准确地把握了他们对新闻内容和好友的关注程度。但现在，他们还认

为有必要把他们对好友的关注程度积极地表现出来，通过访问好友的时间轴或者对他们的贴文做出回应，以便看到好友的更多信息。脸谱网本身就可以大幅增加用户收到的反馈信息，例如，让用户知道贴文中含有哪些内容可能导致好友的信息被屏蔽，哪些贴文可能引起用户的兴趣、让其受到启发，甚至还会知道他是如何以及何时受到情绪感染的影响的。

鉴于人们每天在脸谱网上分享海量信息，科研人员自然希望用这个信息平台做实验，研究人们的心理以及社交网络的效果。在面对面交流时必定会产生的交际效果，例如情绪和情感在交际双方之间的传递，多大程度上会出现在网上交流中呢？研究人员对这个问题尤其感兴趣。脸谱网和康奈尔大学的研究人员修改了脸谱网“动态信息”的算法，通过增加或减少利用积极或消极词语表达情感的贴文数量，了解网站上是否存在情绪感染的现象（答案是肯定的）。不过，在他们公开研究成果之后，这项研究激起了人们的愤怒情绪。脸谱网怎么可以操纵我们的感情呢？其实，媒体和营销人员一直在操纵我们的感情，他们把精心挑选的信息呈现在我们的眼前，对我们产生明显的影响。这也正是希腊悲剧、电视购物和大多数热门电视节目的本质。

如果研究人员告诉大家，他们通过修改算法显示或隐藏的都是讨论是否会下雨的贴文，人们或许不会提出抗议。但事实上，同脸谱网的另一项研究一样，研究人员进行的是富有争议性的情绪感染研究。他们决定开展一个“自然”实验，通过分析各个城市的天气状况，了解情绪是如何蔓延的。为什么选择天气状况呢？这是因为他们发现人们在下雨天更倾向于使用消极词语。研究人员自然也清楚，人的情绪是不可能影响天气的。不过，他们在分析用户状态更新信息里的词语使用情况时发现，下雨对人的情绪影响可以通过社交网络从一座城市蔓延至其他城市，即使那些正沐浴在和煦阳光下的好友在看到贴文时，情绪也会发生变化。这些对脸谱网上情绪感染的研究表明，我们极易受到互联网社交网络的影响。

我是科学研究的拥趸，相信实验可以揭开真相。有人对这项研究提出了批评，认为脸谱网应该在研究开始之前告知用户。这种所谓“知情同意”的方法，在实验数量不多且规模和涉及范围较小时是行得通的。研究人员可以和实验对象坐到一起，让他们知道实验的潜在风险和参与实验所能得到的奖励。在脸谱网的这项研究中，所有人在所有时间里都是在

线实验的一分子，因此，知情同意的概念必须进行修改。在询问登录网站的人是否愿意接受数据收集时，仅仅要求他点击“是”这个按钮还不够。比如，欧盟还要求网站获得cookie（网站为了辨别用户身份，进行跟踪而储存在用户本地终端上的数据）权限。如果拒绝cookie，就会导致网站和手机的某些功能（包括个性化服务）无法使用，因此大多数人都会毫不犹豫地接受。大多数人还会不假思索地接受软件和网站的服务条款（例如苹果的44页超厚协议书）。不要说了解这些服务条款的具体内容，我们很可能一个字也不会看。对于大多数人而言，无论他们怎么努力，也不可能看懂实验规程的详细解释（例如，对脸谱网用于生成“动态信息”的算法的详细解释）。这种情况也无法满足知情同意的出于善意但已经不合时宜的标准。

更糟糕的是，通知用户的做法会危及实验本身。如果实验对象知道研究人员正在调查某个问题（例如，“用户发在脸谱网上的贴文中的情绪因素会对他的好友产生什么影响？”），那么他在脸谱网上的活动几乎百分百会发生变化，但是研究人员无法查明这些变化的原因。比如，在调查情绪感染的研究中，用户可能会更热衷于发现涉及情感的内容，同情性回复也可能会增加。为了不让研究人员发现隐私信息，她还可能会检查自己发表的评论。

因此，我们应该坚持要求研究人员以易于理解的方式将相关实验结果告知实验对象，并将研究成果的价值报告给企业和公众。为了增加透明性，企业至少需要在网站上对已经完成的实验做出说明。如果实验非常复杂，尤其是涉及用户的社交网络时，企业可以选用更有效的方法。假设实验期间“动态信息”被修改的用户第一次不是从新闻中获知情绪感染研究的，而是通过脸谱网直接发送给他们的信息了解到这项研究以及他们在其中所起的作用。这条信息可能包含一篇没有出现在用户“动态信息”中的贴文，并且告诉用户，如果他被分配到“对照组”，他就应该看到过这条贴文了。在理想状况下，如果用户感兴趣，可以申请加入“实验组”，对他们当前的“动态信息”进行修改，并看到实时效果。通过这种告知方式，用户可以很容易地了解数据服务商的选择以及社交网络对他们的影响。同时，用户也有机会表示他们未来是否有兴趣参加类似的研究。

此外，社交图谱及其效果研究还能让你受益。例如，脸谱网可能会提醒你注意，粗略阅读了某位好友的贴文之后，你深受启发，思路豁然

开朗；而阅读另一位好友的贴文，哪怕只是粗略地浏览一遍，你的工作效率也会急剧降低。脸谱网掌握着足够多的数据，熟知你的情绪与工作效率的波动情况，因此在为你选择显示内容时会适当进行调整，以帮助你实现当天的目标。你可以安装某个应用程序，并随时告诉这个程序你在干什么、效果怎么样，从而把你的工作效率和感觉记录下来。或者，你可以戴上Fitbit记录器、苹果手表等活动跟踪设备，查看生命体征的常规读数。又或者，你可以授权脸谱网使用你的手机或笔记本电脑上的摄像头，帮助你找出因为“浪费”太多时间阅读好友贴文，导致你由满面春风变成一脸寒霜的具体时间。在获得这些数据后，脸谱网会对你有所回报，比如建议你增加或减少与某人的线上或线下相处时间。

在评估经济运行状况、军事行动等重要问题所导致的“国民情绪”时，如果可以观察并测量观点与态度的扩散状况，就可以为评估工作提供丰富的背景资料。情绪感染研究同样可以根据社会规范的变化情况为法律修订工作提供信息支持。2015年6月，美国最高法院宣布同性恋者结婚合法。随后，脸谱网为用户提供了一个可选功能，可以在个人资料照片上添加彩虹镜，以庆祝这一喜讯。次日登录脸谱网后，我惊喜地发现，我的“动态信息”中有超过半数的人都使用了彩虹镜功能。但是，在脸谱网的所有用户中，总共只有约3%的人使用了这个功能。我苦苦思索，为什么会出现这种状况。两年前，一些脸谱网用户为了支持婚姻平权运动，将个人资料照片改成了红色“等号”标志的图案。我查阅了关于这件事的一个早期研究，发现很多用户是在看到若干好友更改个人资料照片之后才采取行动的。使用这个标志的好友数量是一个重要因素，但是个体对他人影响力的敏感程度同样重要。

一般而言，用户几乎没有办法了解脸谱网向他们显示好友信息时所采用的排序方法。也许，从脸谱网对双方共同关注内容的评估结果、一项新的广告功能，对点赞和评论数量较多的好友（这些好友创建的内容更多）进行高亮显示、某个A/B测试，或者类似于彩虹镜等功能的使用，我们可以看出一些蛛丝马迹。此外，好友访问某个用户的时间轴时，他们看到用户的好友信息与用户本人看到的不同。是不是脸谱网认为来访者对用户的这些好友更感兴趣呢？脸谱网没有告诉我们答案。就连我们能看到的好友，我们也无法控制其信息显示的先后顺序，更不用说显示给来访者看的好友信息了。

如果脸谱网希望影响一个国家的政治，它就会把观点“更合意”的贴

文排在前列。哈佛大学法学与计算机教授乔纳森·齐特林（Jonathan Zittrain）指出，脸谱网从引导人们参加2010年国会选举的投票活动时起，就已经开始了“民事工程”方面的实验。当时，几乎所有达到投票年龄的美国脸谱网用户都能看到一条提醒他们去投票站投票的广告。一组用户收到若干条动员投票的“社会性”信息，其中列出了已投票好友的姓名和照片。另一组人数较少的用户收到了一条“提示性”信息，同样提醒他们当天是选举日，但没有提及他们的好友投票情况。脸谱网将这两个实验组与没有收到脸谱网选举信息的对照组进行了比较，从三个方面测量了这两类信息的效果：第一，点击广告按钮，搜索本地投票站所在位置的用户人数；第二，点击按钮通知好友自己已经投票的用户人数；第三，通过对比州投票记录中投票者的年龄、出生日期和居住地，“确认”已经投票的用户人数。研究人员宣称，那条社会性信息促使在选举日当天前往投票站的人数增加了34万。齐特林认真地审视了这些统计数据，并且提出了一个非常重要的问题：如果马克·扎克伯格利用他自身的影响力（以及脸谱网算法的影响力），支持他青睐的候选人，向最有可能为他们投上一票的用户推送一条最有效的投票动员信息，我们有办法阻止他吗？

今天是选举日
这是什么？关闭
在“美国政治”页面上找到你的投票位置，点击“我已投票”按钮，把你已经参与投票的信息分享给好友。
我已投票
投票
0 1 1 3 7 6 6 4
已经参与投票的脸谱网用户人数

今天是选举日
这是什么？关闭
在“美国政治”页面上找到你的投票位置，点击“我已投票”按钮，把你已经参与投票的信息分享给好友。
我已投票
投票
0 1 1 3 7 6 6 4
已经参与投票的脸谱网用户人数
安德雷斯 · 韦思岸、盖姆 · 迪亚斯等 25 名好友已经参与投票。

图3–2 2010年选举日当天，脸谱网用户可以看到的“社会性”和“提示性”投票动员信息
资料来源：转载自罗伯特·M·邦德（Robert M. Bond）、克里斯托弗·J·法里斯（Christopher J. Fariss）、杰森·J·琼斯（Jason J. Jones）、亚当· D·I·克拉默（Adam D. I. Kramer）、卡梅隆·马洛（Cameron Marlow）、杰米·E·赛托（Jaime E. Settle）和詹姆斯·H·富勒（James H. Fowler）发表于《自然》第489卷（2012年9月13日）的论文，题为“一个6 100万人参与的社会影响力与

政治动员的实验”。

齐特林建议法律应该让这类政治影响力合法化。但是，姑且不说这类政治影响力的大小难以确定，直邮、预录电话以及定向电视广告等方式还没有合法化呢。法院认为，这些行为一旦合法化就会损害言论自由。我也赞同这个观点。遏制交流或者引入审查制度都无济于事，我们必须要求数据服务商为我们提供工具，让我们知道它们是如何利用我们的交际数据，将它们选择的信息和推荐内容推送给我们的。

信任的价值

不要把信任托付给金钱，而要把金钱托付给值得信任的人。

——奥利弗·温德尔·霍姆斯（Oliver Wendell Holmes）

你愿意搭陌生人的顺风车吗？你愿意待在不相干的人家里吗？你会借1 000美元给素未谋面的人吗？你愿意让不熟悉的人帮你去汽修店取车或者去幼儿园接孩子吗？每个人在回答这些问题时，都会考虑信任这个核心因素。信任的概念非常复杂，难以定义，也不容易测量。但是，社交图谱可以帮助我们解决这个难题。

信任某人，就是你认为可以根据他以往的行为预测出他将如何对待你。通常，只有你认为在未来的交往中某人的行为举止将对你有利，把你的利益放在心上，你才会说你信任他。名望可以为信任奠定部分基础。所谓名望，就是一个人过去的行为与他在特定领域里的专业知识的综合。名望是人或者节点的一个属性，而信任则是两个人之间关系（用社交图谱的语言表述，就是连接两个节点的棱）的一个属性。

信任不一定都是对等的。你对某人非常信任，但是他可能根本不信任你。数字交互可以反映人们之间相互信任的程度。某个机构通过分析人们交流活动的规律以及电子邮件和聊天的内容，推测出谁信任谁、为什么会产生信任等信息。信任还可以通过社交图谱实现高效传播。不考虑某人的口碑，仅凭我们自己的直接了解才去信任某人的情况十分少见。如果我信任艾伦，艾伦又告诉我她信任马克，我就会选择信任马克，除非马克的行为举止让我发现他不值得信任。如果我对马克的信任减弱，就会影响到我对艾伦的信任，至少在可信任人选推荐方面，甚至

还有其他方面，我对艾伦的信任会减弱。即使我没有直接了解过马克的行为举止，随着越来越多我信任的人向我“保证”马克值得信任，我对他的信任程度也会不断增加。

算法可以增强这些信任链条，增加透明性，为核实、确认双方的身份和名望提供一条新的快捷通道。在eBay（易贝网）、淘宝、空中食宿和优步等电子商务平台上，通常用户彼此并不认识，也没有办法通过共同朋友打听其他人的名望。数据服务商必须根据它们能够获取的数据（或者说服人们向它们提供数据）建立信任模型。卖家、旅馆老板和驾驶员需要经常使用某种平台，必然会在平台上留下大量数据痕迹。而交易另一方的买家、住客和乘客可能只会接受一次服务（或者在接受一次服务之后就变换身份）。为了建立信任生态系统，空中食宿利用用户在网站上创建的数据（例如用户搜索、评分、评论、交流记录等反馈信息）与外部数据，核实用户身份，评估他们的可信度。

真正的透明性包括向用户透露评论者与评论对象之间的关系，例如，用户发表的以及关于该用户的所有已发表的评论清单。这些详细信息有助于人们评估每条评论与自己的相关性。例如，Yelp生活服务点评网站允许用户查看评论者打分最高的几个地方。地理位置分布可以表明这个人点评的是他家附近的地方还是离家较远的地方（有的地方他可能根本没去过）。Yelp生活服务点评网站通过采集地理位置以及其他数据，评估每一名评论者的“信任等级”，以决定他的评论在网站上的显示位置。

不过，我认为Yelp生活服务点评网站必须进一步增加透明性，才能帮助用户判断他们是否可以信任某个评价。如果某家饭店使用的是Yelp生活服务点评网站的SeatMe订座系统，Yelp生活服务点评网站就可以确定评论者是否真的去过这家饭店。亚马逊会为“确认已购买”的产品评论添加相应的标签，Yelp生活服务点评网站为什么不向亚马逊学习，设计一个类似的“确认已造访”标签呢？人们发现，一些所谓的名望管理公司在Yelp生活服务点评网站上帮助它们的客户发表四星或五星的虚假评论。它们的这种造假行为情有可原，据哈佛商学院的副教授迈克尔·卢卡（Michael Luca）介绍，商家在Yelp生活服务点评网站的评级每增加一颗星，它的商业收益就会增加5%～9%。

美团网是高朋团购网（Groupon）与Yelp生活服务点评网络的中国版混合体。这家自称月用户超过2亿人的网站通过分析海量数据集，评

估用户反馈意见的可信任程度。该网站推出的商家优惠券只要被使用，就可以确认评论者确实去过那里，并且购买了他评价的服务或产品。不过，由于美团网拥有包括阿里巴巴和腾讯在内的重要支持者，它们还可以做得更好。阿里巴巴利用支付宝应用程序采集的交易记录，可以让顾客看到商家的信誉度。在中国，商家的信誉度非常重要。为了方便人们通过微信程序购物，腾讯公司允许用户将银行账户和信用卡关联到微信账户上。因此，腾讯不仅可以掌握用户的信息交流规律，还可以收集他们的交易数据。在有了这些数据之后，美团网可以对人们的点评进行评分和排序操作，同时过滤可能的虚假点评。但是，美团点评网从未告诉大家，它们是依据哪些数据决定将某些评论升到顶部、沉到底部或隐藏起来的。如果公开评论者与评论的可信度评估方法，Yelp生活服务点评网站、美团网等企业就可以向用户提供更优质的服务，帮助他们选择更好的消费场所。数据服务商也可以为人们提供工具，把信任变成一种“可搜索”的商品。

数据服务商应该为用户提供一个可以方便地打开和关闭个性化服务的开关。脸谱网有一个隐藏的开关，可以帮助用户以两种方式对“动态信息”里的贴文进行排序。第一种是“最新消息”设置，即按时间排序；第二种是“头条新闻”设置，这是脸谱网算法大显身手的地方。但是，脸谱网不仅应该让用户易于发现这些功能，还应该提供更多的排序选择。大多数用户可能并不清楚排序算法的具体工作原理，但他们照样可以尝试不同的设置，并针对特定情况选择自己喜欢的排序方式。算法何时令他们满意或不满意，最终的评判权在用户手中。举个例子。如果你想搜索好友谈论他们如何在旧金山的几家饭馆大快朵颐的贴文，那么你希望排在前列的贴文作者是一位公认的“美食家”，还是一位酷爱运动的家伙呢？“美食家”在最近的状态更新里展示的精致食品赢得了你（还有其他人）的称赞，而那位运动迷说到水煮花生时都会眉飞色舞，不过他的幽默感也赢得了许多称赞，你到底会如何选择呢？时间与相关性排序只能解决这样的问题。

网上零售商知道顾客有时希望对商品按照价格排序，有时又希望按照评价排序。旅行搜索与酒店预订网站允许用户按照费用、飞行时长、出发与到达时间、转机次数和具体航线排序。亚当·戈尔茨坦（Adam Goldstein）与红迪网的史蒂夫·霍夫曼（Steve Huffman）联合创建的嬉芒网设置了一个“痛苦”的功能，可以为机票价格、转机次数、飞行时间赋予权重，帮助用户在综合考虑多个因素的基础上预订机票。（Google

Flights在线预订网站后来也采取了类似的办法。）利用算法解决决策活动固有的平衡问题是一个不错的办法，但是让用户自主决定各个因素的权重可以取得更好的效果。差旅管理公司CWT对1 500万宗业务和7 000个调查展开了分析，试图发现并量化旅行途中紧张因素给人们造成的损失，包括在工作时间与睡眠时间方面的损失。或许，你会为红眼航班赋予一美元的价值。令人奇怪的是，其他数据服务商并没有为顾客提供这种主动性，尽管它有助于实现一种双赢局面。用户通过改变权重和了解最终哪种综合考虑促使自己做出决定，对自己的偏好有了进一步认识；数据服务商也得到了数据，可以为个人和一般用户提供更优质的推荐意见。用户自主排序与赋予权重的做法不仅可以应用于电子商务领域，还应该推广至社交网络平台。

有了更多的排序选择之后，社交网络用户就可以发现他人生活的规律性。在脸谱网上查询某事时，如果条件足够具体，查询结果就可能会直接指向你，并且样本容量为1。如果你的叔叔正在寻找阿姆斯特丹最受欢迎的地方，那么你希望脸谱网告诉你的叔叔，所有好友介绍他们心仪“咖啡店”的贴文都得到了你的热情称赞吗？

很多机构在交易之前或交易过程中，都越来越倾向于通过社交图谱数据来评估对方的信誉度。几年前，为10%的美国家庭提供保险产品和服务的好事达保险公司（Allstate）提出过一个假设：如果有人有过骗保的经历，那么他们在社交网络中采取欺骗行为的可能性就会增加。这个假设是对同质性（价值观相似的人）的一个巧妙应用。在保险行业，骗保的做法更有可能在好友中蔓延。好事达保险公司每年都会收到数百万个理赔申请，不可能全部展开深入调查。过去，好事达保险公司只能依据一些粗略的原则，例如客户居住地周边区域骗保行为的比例是否较高。如果好事达保险公司可以取得客户社交图谱上邻近节点的数据，负责理赔的工作人员就可以利用这些数据筛选出需要详细核实的理赔申请，防止被骗。

财产保险公司经营的主要是“脱机”业务，因此好事达保险公司希望寻找联网数据源，并向数据中间商RapLeaf公司寻求帮助。数据采集初创公司RapLeaf通过购买的方式，收集了大量电子邮件和社交网络数据，其中大部分是脸谱网数据，包括用户好友名单。这些数据的来源是脸谱网用户同意接入他们脸谱网账户的应用程序（这些用户绝对想不到这些程序还可以发挥其他作用）。首先，数据采集初创公司RapLeaf利

用数据挖掘程序，找出哪些网上身份属于同一个用户所有。然后，它借助公司收集到的脸谱网数据，为好事达保险公司提供人们社交活动的相关信息。数据库可以帮助好事达发现有哪些客户的好友也是好事达的客户，然后依据客户好友的理赔记录，确定该客户的理赔申请应该接受哪种层次的调查。在《华尔街日报》揭露了好事达通过不同来源收集并漫不经心地公布部分客户个人数据的行径之后，脸谱网就禁止了这家公司使用它的用户数据。

当然，脸谱网也正在想办法利用社交图谱数据赚钱。2010年，它从社交网站Friendster得到了一项专利，从此脸谱网用户也可以利用社交图谱数据过滤与其他人相关的内容了。不过，在脸谱网于2015年完成一次改版之后，人们发现这项一度占据媒体头条的专利一心追求的其实是金钱。专利证书上有这样一段话：

> 个人申请贷款时，贷款方会查验这个人的社交网络中全部联系人的信誉等级。如果这些人的平均信誉等级达到最低信誉标准，贷款方就会继续处理他的贷款申请。否则，他的贷款申请将被拒绝。

如果你和你的脸谱网好友在现实世界中的唯一交集就是你们曾供职于同一家公司，或者偶尔凑在一起进行一场篮球比赛，或者他是你的远房亲戚，那么你是否会因为他出现在你的好友名单里就考虑与他共担生意风险？我认为，如果我们的个人声望与那个人在社交网络里的声望明显“耦合”，才更有利于开展合作。其中的基本思路是这样的：我信任我的好友丹尼尔·卡尼曼（Daniel Kahneman），因为他获得的诺贝尔经济学奖，以及其他成就都会让他值得信任。或许我愿意把我的一半声望与他的声望绑在一起，也就是说我为我们的“声望耦合信任系数”赋值0.5。这就意味着，如果卡尼曼的声望评分增加1个单位，我的声望评分就会随之增加0.5个单位。相反，如果出于某种原因，卡尼曼的声望评分下降1个单位，我的声望评分也会随之下降0.5个单位。信任系数将帮助我管理个人身份的一个方面（良师益友、奇思妙想以及他们对我的影响），其精细程度远胜于二元对立选择这个典型办法，后者只能告诉我们某个人是不是我们的好友。

如果信任系数向其他人公开，我肯定需要考虑自己做出的那些选择会传递出什么信息。我也许会把我的所有声望全部绑定到像卡尼曼那样的“蓝筹股”人物身上，但是，由于卡尼曼已经是声名显赫的人物，他的

声望评分不可能继续大幅提升，所以我的声望评分也不大可能大幅上升。因此，如果我的目标是提升自己的声望评分，我肯定会去寻找那些“潜力股”人物。

在所谓的“点对点保险”业务（peer-to-peer insurance）中担任保险中介的德国初创公司Friendsurance在经营模型中引入了一个类似声望耦合的机制。在制订Friendsurance保险计划时，两个人（或多个人）表示，如果有一方报告投保财产受损或被盗，另一方就会向他支付定额钱款（比如30欧元）。由于每位好友的承诺可以帮助投保人承担低保费保单所要求的高免赔额，因此在保险范围不变的情况下，投保人可以降低保费。让好友为投保人的索赔埋单，还可以减少索赔金额。相较于保险公司，人们更不愿意欺骗一群好友的钱，或者是因为他们不希望好友知道他们企图骗保，或者是因为他们不希望好友掏腰包。在一定程度上，投保人向他们的好友保证自己的索赔申请没有掺假，在保险公司支付扣除保险中介公司Friendsurance免赔额后的剩余款项时，他们的好友则利用自己的钱包保证投保人索赔的真实性。从本质上讲，Friendsurance保险中介公司将投保人风险评估的部分工作转交给了其他投保人。如果所有人都知道每过三个月，道格就会丢一部智能手机，他的堂兄弟、堂姐妹中会有人邀请他加入这样的保险计划吗？

一旦引入了信任系数和声望耦合机制，就会对用户的社交图谱产生显著影响。你对他人的信任程度是固定不变还是有所变化的呢？在现实世界里，你对某个人的信任度可能在你们相处过程中受损，甚至枯竭，而且你对所有好友的总体信任程度也不是固定不变的。如果我对我弟弟的信任程度增加，并不需要减少我对其他人的信任程度。信任程度增加，也不会导致信任贬值。信任与金钱不同，流通中的货币数量越多，单位货币的价值就越低。采用信任系数的数据服务商可能想对信任度进行人为限制，例如某些交友网站规定会员每天只能发出一条信息。以领英网为例，它不允许用户在某个专门知识领域为其他人点赞，但它可能会给你100个“信任积分”，让你公开分发给你的联系人。这种做法有助于让你的利益与这家数据服务商的利益保持一致：在你分派、管理你对好友的信任度时，网站可以收集并处理你的信任积分分配与信任程度随时间以及他人的行为与声望发生变化的数据，帮助你做出何时应该向何人寻求建议的决定。

数据服务商也可以让你看到其他用户的信任分数分配与信任程度随

时间变化的情况，让你了解某人是否有不断调整联系人信任积分分配方案的习惯，并据此判断这个人是否有声望投机之举。不喜欢这种投机行为的人也许会选择不再信任他。如果你的信任积分来源被公之于众，而且其他用户发现你获得的信任来自社交网络中非常弱的联系，他们就可以自行决定是否介意这件事。有人向你发出“信任请求”，这对你来说也非常重要，因为它表明这个人认为你值得信任，想把他的声望与你的声望绑在一起。信任积分市场的操作方式是否应该与股票市场有几分相似呢？如果你买入微软的股票，那么从本质上看，你是想把自己的财产与微软的财富绑到一起。至于你是否可以购买微软的股票，微软并没有发言权。

信任系数将为人们的决策活动提供有价值的数据，并且为在线社交行为提供规范。如果你信任某人，并且想把自己的声望与他的声望绑在一起，你就有可能更关注他的行为。

建设积极的决策环境

我们经常会征求亲朋好友的意见，认真考虑周围人的是非标准。随着社交数据汹涌来袭，我们的人脉关系（包括交际对象与交流方式）可能会暴露得一览无余。我们是否可以利用社交图谱数据，来改善我们的决策呢？我们的目标不能仅限于提高营销活动的响应率。

通常，脸谱网、领英网等专注于社交图谱数据的数据服务商在数据采集方面占有优势，因为它们提供的平台鼓励人们毫不隐讳地公开自己的社交活动与历史记录方面的信息。但是，这些平台也有短板，例如，我们不可能将个人生活完完整整地搬到脸谱网上。因此，脸谱网不断探索，希望为世界范围内的社交图谱构建更有效的模型。脸谱网的一些插件可以帮助用户为其他网站上的内容点赞，并通过同一次鼠标点击，将这个点赞行为分享到脸谱网上。利用这些插件里留存的cookie，脸谱网可以跟踪用户在互联网上的浏览行为。

脸谱网还可以通过人们共用计算机或移动设备的行为，将他们关联在一起，即便这种共用设备的行为只发生过一次。你登录设备之后，脸谱网就会为设备创建“指纹”，这与BioCatch公司根据键盘使用和鼠标运动规律创建“操作指纹”用于鉴定用户身份的做法如出一辙。脸谱网的设

备指纹是根据多个数据源创建的，包括操作系统的语言设定、已安装应用程序的清单和用户的联系人名单（如果用户同意脸谱网获取这些数据）。脸谱网创建设备指纹的主要用途是保护用户的账户安全。

当然，如果两个用户至少有一次使用同一台设备登录脸谱网（或许是因为他们住在同一个屋檐下），脸谱网就会利用设备指纹技术记录下相关数据。由于很多购买决策都是以家庭为单位完成的，因此这些数据可以用于改进社交图谱测绘与广告活动设计等工作。但是，如果用户决定不把社交图谱中的部分内容分享给这家数据公司，那么，脸谱网能让用户得到多少好处呢？

数据服务商必须努力实现透明性和主动性这两个目标，并向用户解释分享数据为什么可以让他们得到好处。我建议引入信任系数，部分原因是让人们明确表示相互之间的信任，以提高自己的声望。有的人值得信任，有的人信任他人，社会要正常运转的话，这两种人都不可或缺。像信任系数这样新颖独特的方法，不仅会为创建数据的人提供回报，还可以将这些数据转化为决策活动的有力辅助工具，进而影响我们的行为。当然，我希望这是一种积极的影响。

决策离不开具体环境。毫无疑问，社会环境将影响我们的决策，但是，物理环境同样不容忽视。我们来自何方，我们在哪个方面可以发挥作用，这些问题决定着我们接下来的发展方向。在每天不同的时刻，我们依据室外天气状况、疲劳或幸福程度，做出不同的决策。这是最简单的决策活动。是否有人在观察我们，当然也会导致我们的行为（以及决策）发生变化。这些观察结果被记录下来之后，还会导致决策环境发生明显变化。

随着地球上各种传感器数量的爆炸式增长，数据服务商有能力为我们指明方向。至少，我们可以抱持这样的希望，让我们拭目以待！

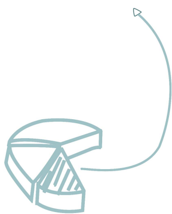


第4章
传感器数据大爆炸的时代

人类社会的传感化

当你的生活被完整地记录下来时，

这对你意味着什么？



有光即可摄影。

——阿尔弗雷德·斯蒂格里茨（Alfred Stieglitz）

街道对面是一栋栋低矮方正的政府办公楼，设计得看不出一点儿新意，在阳光的照射下，纯米色的墙体反射出炫目的光。办公楼前面，一名警察眯着眼睛，看着手里的摄像机，耸耸肩，然后摇了摇头。

男子：你看，我只是在公共场所拍摄视频。而且，我干什么不关你们的事啊。

警官甲：你是说这不关我们的事？

男子：是啊。你们扣留我的话，不是没事找事吗？

这时候，另一名警察掏出便笺薄，宣布这名男子被拘留了。警察将男子按倒在地，确保他没有携带武器。搜身之后，警察对这名男子提出

了警告。

警官乙：告诉你，别让我们为难。我们必须知道你的姓名，确认你来这里不是要谋杀我们。

男子：拍摄你们是违法行为吗？这样做违法吗？

警察盯着摄像机，沉默了很长时间。太阳镜挡住了他的眼睛，无法看清他到底在看什么，但他似乎正在盯着摄像机看。他抿着嘴，咬紧牙，紧锁的眉头上有深深的皱纹。最后，他回答道：“不。”

在当天拍摄的另一段视频中，阳光依然灿烂，但是这一次在阳光下反射耀眼光芒的是警用巡逻车的引擎盖。从车载记录仪中可以看到，这名警察正在耐心地等着交通灯变绿，然后走到一辆正要离开停车场的货车旁边。他先向无线调度台报告自己的位置，再命令货车司机摇下车窗。这名警察让司机靠边停车，因为他的车牌照有一部分模糊不清。

司机立刻告诉警察，他以前拿的是营运车驾照，但是被吊销了。他知道，一旦警察通过无线电将他的车牌号通知调度台，他就会惹上麻烦。果然，调度员说这名司机的驾照无效，于是警察逮捕了这名司机，罪名是在驾照被吊销的情况下驾车。这一切，显然与杰夫·格雷（Jeff Gray）的计划有点儿出入。

格雷是一个市民群体的成员，该群体的目的是在公开场所利用视频记录下政府工作人员的行为，然后发布到一个名叫“拍照不违法”的网站上。当天早些时候，他实施过“第一修正案赋予的监督权”，在奥兰多警察局外面用摄像机记录警察进出警察局的情况，看警察是否会质疑他这样做的权利。没想到，一名警察因此一直跟踪他，并以轻微违反交通规则的罪名逮捕了他。

格雷没有拍摄那次路检以及他的被捕过程，不过，警车上的行车记录仪记录下了这一切。从当时的音频记录可以知道，这名警察先从一般广播频道切换至车辆对车辆的通信频道，然后呼叫在警察局外面隔着街道警告格雷的那名警察。

警察乙：喂，你是要我直接去BRC（违法记录办公室）还是做一些特殊准备？

警察甲：（声音含糊不清）

警察乙：你说什么？

警察甲：弗兰基正在去那里的路上。我们准备找警督，看看还需要做点儿什么。你现在没有在录音吧？

警察乙：你说什么？

警察甲：你现在正在录音吗？

警察乙：哦，我现在是“同车”模式，不过我的麦克风是开着的。

“阳光州”佛罗里达是美国各州中公共记录相关法律最健全的一个州，这对在法庭上质疑警察逮捕自己是否合法的格雷来说比较有利，他有很多为自己辩护的理由。州检察官办公室在业务通讯里告诫警察，只在从污损牌照上无法辨认出车牌号时，才可以命令驾驶员靠路边停车。美国机动车管理局备忘录指出，如果被取消营运车驾照的人驾驶非营运车，就不能逮捕他。他还可以要求法庭播放警车车载记录仪录下的内容，证明他的车牌号是看得清的，而且他还听到警察在讨论是否将他直接带到违法记录办公室或者对他实施“特殊照顾”。

后来，奥兰多的警察们才知道，由于传感器越来越小，存储成本越来越低，人们的一言一行都可以被记录在案。情况已经发生了彻底变化，人类生活的默认状态已经由“未记录”变为“已记录”。

想一想，有多少联网摄像头正在记录你的日常行为。在办公室里、商场里、自动柜员机上、公共交通工具中、城市街道上和汽车仪表盘上，都安装有监控摄像头。你也可能在家里安装了摄像头，监控贵重物品和你的孩子。摄像头几乎无处不在，它们不仅可以记录犯罪行为，还可以发挥震慑作用，阻止人们实施犯罪行为。2011年，根据对英国某郡摄像头全面调查的结果，人们估计仅英国一国的监控摄像头就有200万个，平均每30个人就有一个摄像头。进一步推算，全世界大约有1亿个摄像头在日夜不停地监控公共场所。不过，这只是智能手机摄像头总数量的1/10，后者的数量多达10亿个。几年后，地球上的联网摄像头的数量就可以达到每人一个的程度了。

接下来，我们看看智能手机上的其他传感器，包括至少一个麦克风，利用卫星信号确定位置的全球定位系统，像罗盘一样根据地磁强度与磁场方向确定方向的磁力计，测量气压与相对高程的气压计，测量手

机旋转速率的陀螺仪，测量手机运动状态的加速仪，温度计，湿度传感器，背景光传感器，以及在接听电话时锁住触摸屏的近距离传感器。也就是说，每部手机有10多个联网传感器，仅手机上的传感器总数就超过了100亿，这个数字还不包括汽车、钟表、恒温器、电表等联网设备上的传感器。

如果技术的发展继续遵循摩尔定律，每18个月计算能力就会翻一番，而且价格保持不变，到2020年，全世界的摄像头总数就可能会达到1万亿个。这与惠普、IBM、博世等公司的预测不谋而合。

在奥兰多警察局外面拍摄视频时，杰夫·格雷使用的是一台常规的摄像机，很容易被发现。他拍摄警察的动机之一就是了解警察是否会质疑他采集这类数据的权利，因此他没想着要把摄像机藏起来。勒令格雷靠路边停车的警察知道警车里有一台带麦克风的车载记录仪正在工作，但他没有告诉格雷这件事。格雷的另一动机是在网上公开他拍摄的内容。警察局并没有公开车载记录仪所拍视频的打算，他们是迫于佛罗里达州“阳光”法律的压力，才将这段视频公之于众的。

车载记录仪拍摄的视频为格雷的自我辩护提供了帮助，但并不是所有法律都赋予公民查看这些记录的权利。大多数政府机构都不会将采集到的数据分享给采集对象。关于现实世界交互活动的记录必然会改变人们对隐私权的预期，但同时，人际交往的环境与条件也在发生变化。因此，我们必须考虑一些重要的问题。虽然同为公开场合的记录，但当下法律对照片与视频的态度却不同于音频，这是为什么呢？人所拥有的记录权为什么会受到传感器类型的影响呢？传感器记录的数据应该专属于传感器的拥有者，还是所有与事件相关的人都有权查看呢？如果某人拍摄的视频内容涉及另一个人，那么决定这些数据如何使用的权利应该归谁呢？如果一言一行都被记录在案，人们会因此改进自己的言行吗？如果记录丢失，该做何解释？

涉及奥兰多警察的另一个案例清楚地彰显了连续记录的复杂性。一名被指控醉酒或吸毒后驾车的人在法庭上提出，警方无法提供记录他被捕过程的视频，从而成功地驳斥了对他的指控。至于到底是车载记录仪坏了，还是警察忘记打开车载记录仪，对法官来说毫无区别，没有视频这个证据，警察的逮捕报告不足为证。

随着传感器的数量日益增多，我们必将做出一定程度的妥协，允许

人们收集并分享与我们的身体、情感以及环境有关的数据，尽管有人可能会将这些数据用于我们不知道的用途。现在，我们必须设置某些条件，以确保这些数据给我们（包括个人与社会）带来的回报超过它们所带来的风险。我认为，在使用传感器记录的数据时，必须遵循透明性和主动性这两个原则。

如何充分挖掘传感器数据的价值

从公众对谷歌眼镜用户的反应，就可以看出传感器应用所造成的某些权利不对等问题。为了做一项社会实验，在近一年的时间里，我几乎一直戴着谷歌眼镜。谷歌眼镜非常显眼，一眼就能辨认出。

在墨西哥城国际机场，一位边境检查员将我带到一边。原来，这个区域严禁拍照，而我却戴着有拍照功能的眼镜。实验戛然而止。尽管我的谷歌眼镜处于关闭状态，只有近视眼镜的功能，但这名检查员根本不听我的解释。我争辩说，大多数智能手机都有和谷歌眼镜差不多的传感器，但我的抗议徒劳无功（唯一的“收获”是我在一间狭小的滞留室里“免费”待了几个小时）。

那名边境检查员对谷歌眼镜的反应并非个例。整天不离身的照相机为什么比整天随身携带的智能手机更令人不安呢？那些看到我戴着谷歌眼镜而心感不安的人，对我拿在手里的手机却可能会处之泰然。我不知道为什么会有这样的差别，是因为谷歌眼镜无须手动操作即可方便地拍摄，还是因为它在拍摄时人们难以察觉呢？其实，稍加留意，任何人都会注意到棱镜显示屏上的反光，除非他没看我的眼睛，而在全神贯注地观察周围的环境。

当我戴着谷歌眼镜与人交谈时，人们似乎也会生气，因为他们认为我正在看眼镜里的显示屏，而没有专心同他们交谈。我决定进行一个非正式的实验，在与人交谈时假装查询对方的个人信息，或者假装眼镜正在根据眼前的图像（包括他们的面貌）自动进行“图片搜索”，而我正在查看搜索结果。结果，和我交谈的那些人都感到不快，因为他们觉得自己在对话中处于不利地位。他们习惯于其他人通过电脑或手机搜索获取他们的信息，当我将谷歌眼镜戴上之后，他们看不清在交谈时我是在看他们，还是在看显示屏。

人们感到不安的另一个原因可能是担心我会公开交谈内容。我可以通过手机发送视频，或者将记录的内容上传到云端，供其他人实时观看。如果谈话的一方将谈话内容透露给第三方，或者将对话内容记录下来，分享给其他感兴趣的人，有什么办法可以阻止他们这样做呢？

如果交谈时有人不愿意摘掉太阳镜，很多人都会对此感到不舒服，因为无法看到对方的眼睛，不能更好地了解对方的感受。谷歌眼镜可能不会遮挡住佩戴者的眼睛，但它仍然会对交谈规范构成挑战。人们对谷歌眼镜褒贬不一，说明他们对于无处不在的传感器，以及要么违背要么强行修改社会规范的情况感到担心。总的来说，这种担心主要体现在以下三个方面。

首先，是信息不对称的问题。人们担心对方会公开自己的相关数据，或者担心形势会改变交互的最终结果。如果交谈的一方可以得到另一方无法得到的信息，就会导致令人无法容忍的权利不平衡问题。人们可能担心信息不对称会对自己不利，遭遇“柠檬问题”（二手车销售商向买方隐瞒信息就是一个经典的柠檬问题）。此外，不能专注于当前情境的问题，使人们担心对方的注意力不在自己身上，不安全感加剧。

其次，是信息扩散的问题。人们担心对方不经自己同意便向他人透露数据，或者在公司里或互联网上散布信息。第三方加入之后，正在交谈的人可能会变换话题，第三方显然要先扮演“听众”的角色。但是，在摄像头加入后，这种行为不会自动发生，因此各组织经常会张贴警告标志。由于摄像头的主人、同事或者伙伴观察并分析自己的行为之后，可能会产生某种后果，因此，个人看到这些警告标志后就会适当约束自己的言行。谷歌眼镜本身就是一个警告标志，而且它会不间断地发出警告，即使它处于关闭状态，也不会让人们感到放松。

再次，是信息持久化的问题。人们担心对方记录之后，会将数据保存在某个地方。在这种情况下，人们担心的是一种不确定性，不知道其他人将以何种方式分析、使用这些数据。既然无法保证这些记录会产生积极的结果，就做好最坏的打算吧。此外，关于谁不经同意即有权记录以及有权记录什么内容的问题，法律规定因地而异。例如，个人或组织有权安装摄像头，记录来访者的活动，但来访者则没有这样的权利。关于将私人所有的摄像头连接到无人机上并从空中观察其交谈活动的做法，人们正在讨论和制定相应的规章制度，不过几年内可能难有结果。

在使用可穿戴设备方面，经验最丰富的人应属多伦多大学的电气工程与计算机学教授史蒂夫·曼恩（Steve Mann）了。他佩戴不同版本的“可穿戴智能眼镜”已经有30多年的时间了。20世纪80年代，在麻省理工学院上学的曼恩是“可穿戴计算机”的发明者之一。从那时起，他就几乎全天候戴着他的“眼镜”，并且做了一个又一个应用实验，包括将眼前发生的一切实时上传到网络上（当时，网络直播还很少）。他还创造了“sousveillance”（逆向监控）一词，很少表示他在以视频和音频的方式记录那些在经营场所安装监控摄像头的组织的活动。监控是自上而下的，而逆向监控则是自下而上的。

曼恩主张利用可穿戴计算机增强个人能力。他认为，人们可能不知道某些数据在几分钟之后或者在遥远的将来，会与他们产生某种关系，但是，通过一台“一直运行”的计算机，就可以捕捉到这些数据。为了证明这个观点，他尝试用可穿戴设备来增强人们的感知与记忆能力，例如，通过放大影像或者以超慢速度重放的方式处理人眼无法实时处理的远方物体和信息。根据曼恩的经验，可穿戴设备可以帮助人们过滤输入数据，例如，屏蔽他们不愿意看到的广告。

尽管这些功能有一定的价值，但我认为，想要充分发挥传感器记录下的数据的价值，就必须分享并挖掘这些数据。在佩戴谷歌眼镜做实验的一年时间里，我拍摄了长达数周的视频。但是，绝大多数视频我都没有回看，而且在我做决策、反思自己的行为时，这些视频也没有起到借鉴作用。我无法通过有效搜索从视频中找到相关内容，更不用说对数据进行挖掘，为接下来的工作提供意见和建议了。我有数据采集工具，但却没有合适的工具帮助我找到与当前情境密切相关的数据，也不能对数据加以分析，找到其中的规律或者预测未来。

在接下来的几年里，人工智能技术不断发展，数据服务商可以通过自动化程序为数据添加标签，上述情况将有所改变。企业将会发现所有数据都有分析和处理的价值，包括顾客在商场里的行走路线、员工的专注程度等，而且数据分析技术的价格将大幅下降，大多数企业都能负担起这笔费用。我们也将越来越多地依靠传感器针对具体情境提供的意见。

20多年前，微软研究院的埃里克·霍尔维茨（Eric Horvitz）与美国国家航空航天局的马修·巴里（Matthew Barry），对时间紧、风险性高的决策活动（例如，地面控制台监控飞行安全时需要完成的决策活动）

如何选择信息显示优化处理的时机和方法这个问题进行了研究。首先，他们根据认知心理学的一项经典研究，假设大多数人可以同时处理不超过7条信息。更糟糕的是，在紧张激烈的环境中（例如紧急情况下），有很多因素导致人们无法保持专注，人脑可以同时处理的信息数可能会降至两条。埃里克为工程师监控航天飞机所建立的早期模型，可以在关键的决策时刻断定哪条信息的预期价值最高，并将该条信息突出显示在工程师的电脑显示屏上。

传感器数据的社会化还可以让人们注意到很容易被忽略的重要事项。假设你愿意将某个对话的数据交给数据服务商分析处理，那么你可以使用Cogi应用程序，将最后15秒钟的对话存储到手机的临时缓冲器中，以便找出对话中你最感兴趣的内容。当听到有意思的内容时，只需轻轻按下按钮，应用程序就开始永久性保存这段音频，直到你按下停止键。否则，缓冲器就会不断覆盖之前存储的内容。如果有若干人同时使用该程序记录某个对话，就可以进行比较，确保存储的正是最有价值的内容。一段时间之后，分析你保存的对话，就可以发现哪位发言者、哪些话或者哪些话题最能吸引你的注意。

关注度与相关性因情境而异。在第1章里，我们举了一个用“模棱两可”的词语——美洲豹（可以指猫科动物、汽车品牌或者计算机操作系统）作为搜索项的例子。从这个例子中我们看出，算法可以根据多个类别为搜索结果排序，并高亮显示最符合你的搜索意图的信息。了解你所处的情境，有助于数据服务商提高数据产品的相关性。例如，假设你站在动物园中，利用手机搜索“美洲豹”。如果应用程序成功定位到你的地理位置，就会将猫科动物美洲豹的相关信息放在搜索结果的前列。如果你站立的位置是动物园的停车场，应用程序通过手机摄像头了解到你前面有一辆最新款的豪华轿车，它就会推测你对这辆车感兴趣，而不是在结束了一天的动物园参观之旅后还希望了解美洲豹某个方面的生活习性。

不过，并非所有与情境相关的搜索项都像这个例子那样清楚无误。如果某人在俱乐部待到深夜之后搜索“茉莉”（jasmine）这个词，他不大可能想要查找茉莉花的相关信息，而更可能是回家路上一家24小时营业的中餐外卖店的地址，也有可能是想查阅成人娱乐网站Livejasmin上的色情服务信息。他是在市区游荡，还是待在自己的卧室里？数据服务商可以通过他近期和当下的地理位置数据，针对性地为他提供搜索结果，

确定他希望前往的目的地。

考虑情境的相关性，从长远来看有助于我们的决策活动。用丹尼尔·卡尼曼的话说，思考宜“慢”不宜“快”。例如，某些银行考虑根据个人的交易记录和当前的情境，为顾客提供“无悔式”服务。如果你在凌晨4点走到拉斯韦加斯赌场的一台自动柜员机前，准备提取1 000美元，机器不会立即吐出现金，而是提示你：“你确定现在要提取这么多的现金吗？在这种情况下按下‘确认’按钮的人大多会后悔”。

如果传感器在你的掌控之中，你就可以设置条件，决定何时将你所处的情境告知数据服务商。但是，在未来10年里记录你生活点点滴滴的一万亿个传感器中，有些是被银行、商场、公司、学校和政府部门控制的。这些机构和组织日益感兴趣的不只是你在特定时间里的位置，还有更多的隐私信息，比如，你和谁在一起，你有什么感受，你在关注什么（而不是“你应该关注什么”）。但是，你的“完整”的情境到底何时起到影响作用，应该由谁说了算？在回答这个基本问题之前，我们必须先了解你无法完全掌控的传感器数据是如何泄露你所在的情境的。

雇用私家侦探的做法过时了！

世界标准时间2000年5月2日凌晨4点，美国政府停止了向美国国防部24颗导航卫星输送噪声的行为，使GPS（全球定位系统）的分辨率精确到米这个数量级，也使一系列个人导航服务成为可能。向普通大众提供这种精度的导航服务，产生了巨大的效益。据估计，仅通过提高经营效率，GPS在2013年就为美国经济多带来了700亿美元的收益。此外，GPS在改善医疗、安全和环境管理方面的贡献仍然有待统计。

得克萨斯大学奥斯汀分校的工程学教授托德·汉弗莱斯（Todd Humphreys）认为，民用GPS还有可能更加精准。在三星公司的资助下，他和同事把普通手机GPS的精确程度提升至1厘米。据汉弗莱斯预测，10年内，我们将会在几乎所有财物上安装微型定位跟踪器；一旦财物丢失，就可以像在网上搜索信息一样，搜寻这些失物的下落。但是，现在的GPS只能让你知道你自己所在的位置，或者让其他人找到你。

重量与钢笔相仿、大小与邮票差不多的定位跟踪器不是科学幻想，

而是已然存在。这种跟踪器由一块硬币大小的电池供电，电力可维持一年时间。之所以能做到又轻又小、高度节能，是因为它们是按照信标的标准设计的。不过，它们不会接收导航卫星的常规信号，而是通过低功耗蓝牙（BLE）协议发射一种独特的识别码，由某些设备在10米范围内接收这些信号。用户的定位跟踪器识别码被发现后，他安装的应用程序就会把时间和地点报告给生产商的中央数据库。有时，用户与信标之间的距离非常近，借助手机上的应用程序，就可以找到信标。但是，大多数生产商为用户提供了一个选择，借助安装有该应用程序的其他手机扩大搜寻范围。据报道，已经有人利用信标寻找丢失的钥匙、追踪失窃财物、在密集的人群中确定爱人的位置。此外，脸谱网等大型数据服务商还向各种组织机构提供免费信标，以普及信标的应用。利用这些信标，脸谱网的应用程序在用户接近某个脸谱网信标时会发出提示，还可以向用户提供与本地相关的内容，记录他的活动。

如果有人悄悄地将蓝牙信标塞进他人的手袋或钱包中，我们有办法阻止他吗？没有！目前，法律并没有禁止通过定位跟踪器追踪他人的行为（但是，如果跟踪者为政府工作，就必须有搜查证，方可实施跟踪行为）。想要保护自己，不让他人通过连接传感器的方法跟踪你，那你只能使用自己的设备。

我所说的“自己的设备”，是指这些设备真正为你所有。目前，隐私权保护运动方兴未艾，便携式蓝牙干扰仪、GPS干扰仪正在占领市场。这些干扰仪可以产生与蓝牙或导航信号频率相同的噪声，使干扰仪周围几米内的所有信号接收装置全部失效。尽管在网上可以找到干扰仪的制造方法，但在包括美国在内的多个国家，制造、销售和使用干扰设备的行为是非法的。不过，人们并没有因此停止这方面的尝试。在一个案例中，一名卡车司机不希望公司在上班期间跟踪自己的活动，就在公司的卡车上安装了一台GPS干扰仪。干扰仪的效果的确不错，但它不仅干扰了公司的GPS，还导致纽瓦克国际机场的空中交通管制系统陷入混乱。（卡车司机试图干扰公司了解他所在位置的努力，为他带来了3万美元的罚单。）

干扰仪不仅会阻止GPS设备发现它所在的位置，也有可能报告不正确的定位数据，欺骗GPS。托德·汉弗莱斯发明的GPS欺骗攻击系统可以向目标接收装置发送虚假的卫星信号。在人或车辆的应用程序激活GPS之后，欺骗攻击系统会故意将他们引往错误的方向。

这些跟踪器、干扰仪和欺骗攻击系统也许会让你觉得胆战心惊，以致再也不想使用任何GPS或蓝牙设备了。但是，即使功能最简单的手机也可能会暴露你的位置。一天之内，手机会不停地切换基站，你的活动因此被记录下来。如果手机是通过WiFi热点接入互联网的，热点提供者就可以精准地了解你在网上的活动。与GPS不同，WiFi的工作环境是室内。现在，很多零售商提供免费WiFi，因为这项服务有助于它们观察你在店内的活动。（WiFi信标与蓝牙信标都可以收集这些数据，两者构成竞争关系。）这是一个非常重要的关系反转。对零售商来说，与其为你上网查询产品信息、顾客评论和竞争对手的价格制造障碍，还不如想办法了解你的确切位置。有了这些数据，零售商甚至可以针对特定位置提供特别报价，无论这个位置是某个摆满商品的销售区，还是你徜徉其间的货架通道。

还有一种完全不同的数据源，可以透露你曾经的位置信息，那就是你拍摄的照片和你被拍到的照片。首先，在网上公开发布的照片大多是用手机拍摄的，而大多数有照相功能的手机都有GPS。照片默认关联的元数据包括照片拍摄地点的经纬度，尽管你可以把这些元数据从你自己拍摄的照片中删除，但对于其他人拍摄的照片，你就无能为力了。每天，人们都会拍摄不计其数的照片，因此你的位置很有可能被记录下来。

地理位置元数据并不是照片携带的唯一信息。照片背景中的著名地标、街道指示牌或者饭店的菜单都有可能透露你的位置，影子的长度可以大致反映当时的时间。算法也在接受视频的训练，即使是分辨率较低的监控摄像头拍摄的不清晰视频，算法也可以根据步态特征，识别人行道上行人的身份，并在他走过一个个街区时实施跟踪。

尽管有人提议戴墨镜、帽子，化妆和使用假胡须，以迷惑监控你的设备。但是，要摆脱人脸识别软件的监控，难度已经越来越大了。我们在第2章讨论过，脸谱网的DeepFace软件可以比对你之前被标签的照片，从中找到你，无论这些照片的光线条件和拍摄角度的变化幅度有多大。你也许会要求好友删除他们给照片添加的标签，但有的标签是机器生成的，你可能根本看不到，又该如何处理呢？此外，还有若干企业正在引入身份验证程序，要求用户公开自己的动态照片或视频。

腾讯公司是这个领域的先行者。腾讯发现一些用户对QQ聊天工具进行了不正当使用。不少卖淫者注册QQ账户用于性交易，他们往往会

用从网络上剽窃来的照片作为个人资料照片，有时还会加上带有法律意味（或者说看似具有合法性）但是根本无用的免责声明，比如“仅做说明之用”。见多识广的人都知道不能相信网络照片，但这种做法十分猖獗，以至于很多人不再信任QQ用户的任何个人资料了。于是，腾讯启动了用视频动态验证用户个人资料的计划。用户需要根据腾讯公司任命的社区经理的要求，分享网络摄像头的视频流，并且按照实时指令完成一系列动作，例如摸右耳、耸左肩等。如果视频中的人脸与用户上传的个人照片相吻合，账户资料就可以“验证通过”。

腾讯的这项计划需要雇用大量人员发出指令和评估视频，而且至今仍然有许多的账户没有验证。随着人脸识别技术取得迅猛发展，机器已经能够胜任这项工作，而且可以全天候工作。支付业务运营商Worldpay开发的PIN码（用户识别上的个人识别密码）输入设备摄像头，可以在顾客向商店的支付终端输入银行卡密码时拍摄一张或多张顾客的面部照片。支付业务运营商Worldpay计划建立并维护面部照片中央数据库，用于升级人脸识别软件，以确认正在使用银行卡的人是否有相应权限。如果照片不吻合，店员就需要在购买行为完成之前用另一种办法核实顾客的身份。为了遏制利用遗失信用卡或卡号在线购物的行为，万事达信用卡同样推出了身份核实计划。信用卡客户上传指纹扫描图与面部照片，就可以建立个人的生物统计档案。需要核实身份时，他们通过手机提交指纹扫描图和“自拍”视频，同注册的生物统计学数据进行比对。如果图片不吻合，支付请求就会被拒绝。在早期试用时，用户需要在视频中做眨眼的动作，以证明自己是活生生的人，以及照片与信用卡信息不是从网上盗来的。其他银行或其他机构也建立了指纹与面部照片数据库，或通过购买的方式取得数据库的使用权，保护客户及其财产安全。如果不愿意个人数据被采集用于这个目的，可能就无法享受到某些财产保护措施。

除了留在触摸板上的指纹以外，还有一些独特的生物标记可以确定人的身份，例如眼睛里虹膜色素沉淀的特征。与指纹不同，虹膜永远不会磨损，这意味着这个生物标记可以使用更长的时间。指纹扫描要求手指触摸扫描仪或其他采集设备，而虹膜可以用摄像头在10米以外进行扫描和识别。卡内基–梅隆大学（位于匹兹堡）的几名生物统计学研究人员利用汽车后视镜的反光，在人进行室内漫步时，成功地拍摄了他的虹膜照片，并与之前储存的虹膜记录进行了比对。若干个国家，包括拥有10亿人口的印度在内，正在（或计划）要求更换国民身份证的公民留存

虹膜扫描图。

不过，在鉴定身份或进行精确定位时，被鉴定人无须露出自己的脸。除了手机以外，汽车是确定你的身份和你所在位置的最佳途径。法律规定，汽车的车牌号必须清晰可辨，杰夫·格雷对这条规定就非常清楚。位于加利福尼亚州利弗莫尔市的Vigilant公司把成千上万个摄像头拍摄的照片加以综合。这些摄像头的安放位置各不相同，有的在停车场里，有的在商店橱窗上，有的在私人住宅的外墙上。将传感器报告给中央数据库的独特识别符进行汇总，就有可能重现某个人过去的活动。Vigilant公司的摄像头拍摄的照片经光学字符识别算法处理后，就可以确定每辆车的牌照。Vigilant公司声称，他们每个月都会向公司的美国数据库添加约1亿个私家车车牌识别码。

除了位置固定的摄像头以外，Vigilant公司还说服大量司机在车里安装车载记录仪，并将数据传输到公司的数据库里。很多这种车载记录仪都安装在“回收”公司调查员驾驶的汽车里，目的是寻找车主未继续偿付购车贷款的汽车。有了这些车载记录仪之后，Vigilant公司大幅增加了数据库的覆盖范围。

Vigilant公司的这种汽车牌照数据库有两种查询方式。你可以查询某个汽车牌照，了解它在何时何地被系统发现；也可以查询在某个地点和时间的汽车牌照。警察经常查询在犯罪地点出现的汽车牌照，目的是寻找嫌疑人和目击者。如果某人的汽车出现在附近，那么他本人也可能在附近。与普通人不同，警察可以查阅各州的汽车数据库，根据汽车牌照找到车主姓名。美国公民自由联盟（ACLU）透露，美国警察局曾宣称，在使用这些数据库时，“没有他们做不到的，只有他们想不到的”。

据报道，Vigilant公司与其他私有汽车牌照数据库的主要客户是执法机构，但从理论上讲，任何人都可以接受Vigilant公司的付费服务。公司的委托人包括希望收回汽车的汽车经销商和希望了解事故详情的保险公司，私家侦探也会使用这些数据库。对你的亲朋好友而言，你的车牌号不是秘密，他只需看看车库即可了解这个信息。如果你怀疑自己的配偶是否真的每晚都在办公室加班，雇用侦探跟踪他的做法已经过时了。最有效的做法是将他的车牌号输入数据库，查询他的汽车去过哪里，以及是否有其他汽车也到过那里。如果你还想知道那些汽车的驾驶者是谁，数据库对汽车经常停放的位置都了如指掌，很可能据此找出驾驶者的居住地点与工作场所。一度成本昂贵，甚至具有危险性的数据收

集工作，现在已经毫无危险可言，费用也大幅下降。

美国一部分州曾试图查封商业性质的汽车牌照数据库，但是人们站在言论自由的立场上对此提出了质疑，使得这项法律至今未在任何州获准通过。在公共场所拍照不违法，将图片存到云端不违法，使用光学字符识别系统也不违法。如何才能保护人们的隐私权呢？我们可以不用车牌号，代之以一种路人与摄像头看不见、可以发射独特加密代码的设备。政府必须每隔一段距离就安装一个接收装置，接收这些设备发出的信号。当然，这些设备有可能被非法入侵，就像汽车可能安装假车牌一样。此外，私营企业也有可能制造传感器，探测这些设备。毕竟，世上没有十全十美的事。

摄像头并不是可以捕捉人、地点和事物独有特征的唯一一种传感器。麦克风（包括手机里的麦克风）可以捕捉周围环境中的噪声，获得足够的数据，根据汽车引擎的震动、车体的抖动以及轮胎发出的声音，识别你驾驶的汽车。

既然音频分析软件可以识别你的汽车，它当然也会知道你是坐在停着的汽车里还是行驶中的汽车里。而且，无论你是否打开GPS，它都不会受到影响。在手机麦克风上接入从语音到文本的转换程序，例如苹果的Siri（语音控制功能）和微软的Cortana（微软小娜），还可以利用周围环境中的噪声分析你所在环境的特征。

大型数据服务商也在生产家用传感器。我的卧室里安装的亚马逊Echo智能家居服务系统就一直处于待机状态，通过7个麦克风监控各种风吹草动。用户指南告诉我，只有听到热词“亚力克斯”，设备才会被激活，开始一字一句地记录下我说的话。比如，根据我的问题或指令，从网络上搜索相关信息，或者根据我的要求，将商品放进我的购物车，形成一个虚拟的“待购买”商品清单。与之类似，微软针对Xbox游戏控制台开发的Kinect系统和三星的智能电视都可以识别语音指令，包括待机状态下根据指令自动“打开”。谷歌的Nest Cam联网安防摄像头带有麦克风，你不在家时，如果房子里出现异常的声音（例如，入侵者的说话声），它就会发出警报。

这些设备是不是无须你特意要求，就会分析你说的每一句话呢？至少对于其中的一些设备而言，情况确实如此。事实上，一家公司的一位研发人员私下里已经亲口承认了这一点。这位研发人员所在公司利用这

些设备收集的数据，尤其是有背景噪声干扰的谈话，改进语音识别软件的效果。在安静的家庭环境中，音频分析可以为每位家庭成员构建清晰的声纹，展现家庭交流的规律（和频繁程度）。要发现外来者的声音，谷歌的Nest Cam必须先学会识别家庭成员的说话声。如果确实如此，谷歌为什么不告知用户他们正在收集数据以改善系统的性能呢？告知用户这个信息，不是更有利于提高透明性吗？毕竟，谷歌不是供暖器或婴儿监护器的生产商，而是数据服务商。

如果企业进一步公开这些音频数据，或许我们可以从中得到一些好处。就我个人而言，我希望可以获得我的所有语音记录，因为这些记录可以发挥神奇的作用，比如，追踪已经被我淡忘的信息，或者发现我的语言表达的有趣规律。为了改善语音识别系统的效果，数据服务商必须编制语音与词语索引。宠物生病时，我可以借助这个索引，找到我说过的每一句和这只宠物有关的话，然后把一周以来宠物的行为举止告诉兽医，为他的诊治提供更丰富的辅助信息。

走出家门，环境中的声音也会透露出我们所在的物理环境。站在人行道上，坐在附近的汽车里，或者从房间里眺望街道，我们听到的警笛声或者汽车喇叭声大不一样。酒杯碰撞的喧嚣声是饭店的明确无误的标志，类似的标志还有裁判吹响的尖利哨声、波浪拍击礁石的声音、铺瓷砖的浴室里响起的回声。如果你正在东京附近的一辆火车上，每个车站播放的独特音乐就会透露出你所在的位置。在打电话时，几乎不可能将所有这些可以清楚显示所处环境的声音线索清除干净。

传感器数据还可以被用来构建社交图谱。假设两个人决定进行秘密幽会。他们知道，分析手机位置数据，就有可能发现他们在当天的同一时间里出现在同一个地点。因此，他们商量好在距离约会地点几个街区的时候关闭手机，给自己加上“隐私屏蔽层”。在惩罚降临的那一天，他们的电话记录变成了呈堂证供，在几个小时里两人手机同时无法追踪，是否能证明这两个人在一起呢？也许可以吧。如果当时两部手机正在朝着同一个目的地移动，这个信号就更加明确了。

社交图谱也有助于寻找人们的地理位置。一名德国逃犯潜入加拿大境内后，十分明智地抛弃了他的手机和SIM卡（客户识别模块）。不过，尽管他不再使用以前的SIM卡和手机号码，但他仍然会用新手机给别人打电话。任何人的电话呼叫图谱都与众不同，国际刑警可以通过寻找突然出现，而且呼叫规律相似的新号码，来抓捕这名逃犯。在他打了

10多个电话之后，国际刑警就标记出这个可能符合条件的号码，并通过定位这部手机，抓获了这名逃犯。

数据服务商分析传感器数据的行为可以让大多数用户获益，但被人窥探却毫不知情的那些人却有可能受到伤害。如何取得平衡，是摆在他们面前的日益迫切的任务。创建于2011年并被谷歌于2014年收购的图片分析创业公司Jetpac，专注于对照片的内容进行识别和归类，希望编制一个可以根据特色进行查询的企业名录。公司利用全世界6 000多座城市的Instagram用户上传到网上的1．5亿张照片，对应用软件进行了训练。其中很多照片带有地理位置标签、井号标签或者标题。如果照片的拍摄地点有很多人涂口红，应用程序就会认为这个地方的人“时尚优雅”。这一类信息可以帮助人们判断某个地点是否适合自己前往。

图片分析创业公司Jetpac认为，利用它的目标识别软件，可以罗列一个全世界的“顶级嬉皮士酒吧”名单。该公司的数据科学家认真研究了照片中留八字胡的人所占的比例，据此估计某个地方的嬉皮士人数。结果他们发现，嬉皮士数量最多的城市似乎都在土耳其，为什么会这样呢？这些数据科学家意识到，土耳其男性对八字胡的热衷程度远高于美国男性。因此，他们需要针对不同的地区设置不同的评估标准，结合地方传统实现数据的“归一化”。这个案例是关于数据分析过程中计算机与人之间基本反馈回路的杰出范例。

图片分析创业公司Jetpac在分析这些Instagram照片时，还发现了一些其他内容，结果导致了更棘手的问题。例如，Jetpac公司发现，仅凭它掌握的数据，就可以罗列出德黑兰市同性恋酒吧的名单。这对通过询问朋友或陌生人以避免找错对象的伊朗人来说，是一项非常棒的服务。但是，如果这份名单落到毛拉们的手中，同性恋群体就会面临极大的危险。不过，虽然Jetpac公司具备这种数据控制能力，又有什么办法可以阻止政府紧随其后呢？

麻省理工学院教授威廉·T·弗里曼（William T. Freeman）带领同事研发的算法，可以侦测到皮肤颜色的像素级的微小变化，进而测量人们的脉搏，包括血液在人体裸露部位的分布情况。弗里曼团队演示了他们的研究成果之后，人们纷至沓来，要求他们提供这套算法。于是，他们将算法发布到网上，允许任何人用于非商业用途。一位扑克牌玩家想利用这个算法来侦测对手玩牌时是否心跳加速，以便判断他是不是在虚张声势。

无独有偶，心脏病专家发现人们心跳的波峰与波谷（即所谓的“PQRST波形”）与指纹或虹膜一样具有独特性，而且篡改的难度更大。万事达、加拿大皇家银行和英国的哈利法克斯银行，在自动柜员机服务、网上银行与非接触式支付业务中，尝试通过一种心电图（ECG）手环来验证客户的身份。这种手环的制造商是位于多伦多市的可穿戴式验证设备公司Bionym，它的设计思路源于测量、记录与验证心律的多项专利。

我们不可能掌控全世界数万亿个传感器收集的全部数据。Jetpac公司的联合创始人、首席技术官皮特·沃登（Pete Warden）明确指出，隐私与安全、隐私与自由言论之间的平衡态势将发生本质性变化，图片识别软件将在其中起到推波助澜的作用。一方面，企业正在寻找更安全的方法保护敏感数据。越来越多的企业要求人们在接受特定服务或隐私保护时，必须使用生物统计数据这把独一无二的“钥匙”。另一方面，我们的社交数据有很多都是通过相同的兴趣，同时参加的某些活动以及相互之间的关系创建的。从源头控制数据收集以保护个别人的隐私权的做法，必然导致很多人的言论自由受到过分的限制。

此外，尽管我们有三种方法可以保护自己的传感器数据不被别有用心的人利用，但是这些方法并不完善。要求使用电子钥匙才可以获取数据的方法不适用于很多公开数据，例如在Instagram上发布的照片。指导我们谨慎分享、使用数据的社会规范，无法阻止不良分子对我们隐私的侵害。这样一来，我们能选择的方法貌似只有法律法规了。但是，从路易斯·布兰代斯侵犯隐私权长达20年之久就可以看出，法律对动态技术革新的反应速度非常慢。另一方面，如果目标是界定宽泛的数据集，例如种族、性别、性取向和伤残状况等，法律体系的相对稳定性就是一大优势。但是，使用这些数据集往往会被视为歧视行为和不受欢迎的行为。

不过，单单增加可选方案的数量是不够的。我们还必须备有工具，可以侦测这些基于社交数据应用的歧视行为。对数据库的每一次查询都是一组数据，经数据服务商收集、分析之后，有助于改善它们的产品与服务。此外，这些数据也可以用来保护我们。这种保护作用在当今社会尤为重要，因为算法不仅可以找到我们身在哪里，还可以窥探我们的内心世界。

加利福尼亚大学旧金山分校的心理学荣誉退休教授保罗·艾克曼（Paul Ekman）一直在研究6种基本情感的生理效应。这6种情感分别是生气、伤心、害怕、蔑视、惊讶和愉快。艾克曼让来自5个不同国家（智利、阿根廷、巴西、日本和美国）的人看这6种情感状态的照片，并观察他们有何反应。他预测文化环境的不同会导致人们的反应各异，但结果证明他错了。这个实验重复了许多次，他发现人们在看照片时都会产生相同的表情：与生气相关的是眉头紧锁，眉毛和嘴角下垂表示看到的是伤心的照片，皱鼻子表示蔑视，与真诚笑容相关的是眼角纹。（礼节性假笑——因为泛美航空公司的乘务人员总是面带这种笑容，因此又被称作“泛美式微笑”——往往只有嘴部有变化。）1978年，艾克曼与他的同事华莱士·弗里森（Wallace V. Friesen）通过总结他们观察到的所有表情，建立“面部表情编码系统”（FACS）。根据FACS，几名机器学习研究人员开发出了人脸识别软件。

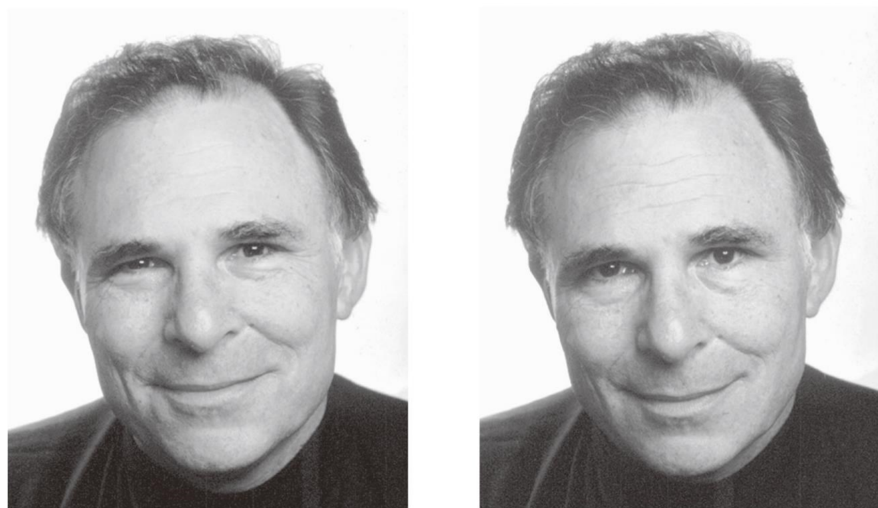


图4–1 真诚微笑（左）与礼节性微笑（右）的对比。人在真正高兴时，他的眼角与嘴角的肌肉都会运动，导致皮肤产生皱纹
资料来源：由保罗·艾克曼博士和保罗·艾克曼有限责任公司提供。

艾克曼假设，情感具有普遍性，因为情感是反映我们的心理状况和彼此关系的真实信号。随着实地研究与实验的进行，他发现每种基本情感还与其他生理指标有关，例如心率、呼吸率、血流量和肌张力等。有时，人们的情感变化非常快，如果不注意，甚至难以捕捉到情感变化的过程。这些“微表情”常常意味着这个人不想表露自己的情感，或者他没

有意识到自己的这种情感。由于这些表情稍纵即逝（持续时间大约只有1/5秒），没有经过专业训练的话，是很难发现的，需要借助回放视频才能看到。

艾克曼曾经在圣迭戈一家名叫Emotient的公司担任顾问，该公司开发出了可以从摄像头记录的原始数据中实时识别情感的软件。2007年，Emotient公司推出的第一个商业应用程序是“笑脸检测程序”，可以安装到索尼数码相机上。当取景框里的人露出笑容时，该程序会立即抢拍。随着Emotient公司的算法不断进步，一台高清照相机就可以监控同处一室的400人），同时“读取”他们脸上的微表情。该公司还打算将这款软件推广到医学疗域，用于捕捉患儿脸上的痛苦表情。事实上，研究表明，在捕捉身体不适的真实信号这个方面，计算机强于人类。Emotient公司与谷歌早期合作开发的一种眼镜应用程序被推销给公司管理者，帮助他们了解员工的精神面貌，以及情感对顾客购买行为（买什么？从谁那儿买？）的影响力。2016年1月，Emotient公司被苹果公司收购。

伦敦的Realeyes公司也引进了艾克曼的研究成果，以评估人们在看到广告视频时的面部表情。广告显示屏可能是某个人的电脑显示器，也可能是安装在公共场所的显示屏。电子产品生产商LG开展的“舞台恐惧症”广告活动就以男厕所为广告背景。厕所的小便池上方装有LG显示器，当有男子来小便时，屏幕上就有一名女子把广告推开，摆出一副能看见男子小便的姿态。据Realeyes面部表情分析摄像头的观察，这些男子的表情由困惑、害怕变成高兴。分析结果甚至表明，有一部分男子在视频开始和结束时都表现出厌恶的情绪。麻省理工学院媒体实验室情感计算小组找到Affectiva公司，合作研发可以帮助孤独症患者解读他人面部表情的情感警报系统，从而与Realeyes公司形成了竞争关系。商业客户请Affectiva公司检测人们对广告视频的情感反应，民意调查公司则用它来统计电视辩论期间政治候选人的支持率情况。

在收集、分析人的情感数据时，除面部表情外还有众多数据来源。人说话的音调、音量（声强）、语音质量、持续时间和语速等也能反映人的情感状况。为了创建情感语音检测系统，一些研究人员从5个说英语的国家（澳大利亚、肯尼亚、印度、新加坡和美国）雇用了100名演员，请他们声情并茂地把一些简单的文本（诸如日期和数字）表达出来。

研究人员认为，这些语音表达过于“装腔作势”，不能成功地训练机

器学习系统实时处理真实的人际对话。一些近期的研究旨在利用客户呼叫中心的海量数据，建立情感图书馆。呼叫中心安排业务代表在接听电话的过程中记录客户的情感状态，在音频记录上添加标签，以创建用于机器学习的数据。人们已经在利用这些添加了恼怒、温和、激烈、中性等标签的语音（甚至包括“啊”、“哦”、“嗯”、“好的”等非常简单的言语），对语音识别系统进行训练。在某些情况下，语音信号与用户满意度调查数据相结合，可以验证情感检测系统的准确性。

云联络中心服务提供商LiveOps公司、人工智能公司Mattersight等利用语音检测软件为客户安排服务代表。如果客户有很重的地方口音，就为他安排一名家在该地区的客服代表，让他们进行更有本地特色的交流吧。如果客户听到呼叫中心的选择菜单后立即做出选择，这可能意味着这名客户十分生气。应该怎么办呢？把他的电话转接给善于处理难题、应对挑剔客户的服务代表。如果客服代表已经竭尽全力，仍没让客户平静下来，而且客户的声音越来越大、越来越尖利，这个呼叫就会升级，交由业务经理处理。除了这些音频数据，LiveOps公司还会针对客户投诉搜索社交媒体和其他数据源，寻找更多的背景资料。有的客户很快就和呼叫中心的客服代表建立了融洽的关系，这次投诉可能就不难处理，而且客服代表还有可能说服这名顾客购买产品或服务。人工智能公司Mattersight宣称可以根据客户的性格类型安排客服代表，为客户提供更有针对性的服务。该公司利用交流记录分析对话内容与方式，把客户分为“开朗、尖刻、严肃、内向”等类型，并把客户的电话转接给善于同这种性格类型的客户打交道的客服代表，以增加顾客的满意度。这种安排的依据是性格类型，而不是呼叫本身的特点。该公司的很多客户都是需要经常与客户交互的企业，例如医疗保健企业、保险公司和电话公司。

算法还给那些缺乏表达能力的人带来了福利。人们常说，父母可以分辨婴儿哭声传递出的情感需求。但总的来说，这种能力不具有科学性，显而易见的原因就是可供父母学习的样本太小。在与周围世界交互的过程中，人和机器为交互数据建立模型的方式存在若干不同之处，样本大小是一个明显的不同点。参与开发谷歌无人驾驶汽车项目、教育领域初创企业优达学城（Udacity）的联合创始人塞巴斯蒂安·特隆（Sebastian Thrun）指出，驾驶员凭借个人经验开车，而谷歌无人驾驶汽车可以从所有无人驾驶汽车犯下的错误中汲取教训，提高驾驶技术。人主要是从自己的成败经历中吸取经验，社交图谱中其他人的成败仅起

到辅助作用。此外，他们还可以征求专家的建议。相比之下，机器不仅可以直接从它们犯下的错误中吸取经验，还可以从其他机器所犯的错误中得到教训。

IBM的迪米特里·坎尼夫斯基（Dimitri Kanevsky）和同事开发的一项专利技术，可以从婴儿的啼哭声和大脑、心脏及肺部活动中采集数据，开展学习。婴儿哭闹的原因有很多，有时是为了引起注意，有时是因为孤独。数据服务商可以帮助父母们更准确地监控孩子的情感状态，并依此做出决策。

将来，除了面部表情、啼哭声的音调和音量以外，应用程序还可以根据其他更微妙的线索探查我们的情感状况。一些活动追踪系统（例如Fitbit记录器、Withings Pulse智能手环、佳明智能手表）可以记录人们的生命体征，包括静态心率和运动心率，这些生命体征可能与某些情感状态有关。血液流经身体时，皮肤上的红色会加深，因此利用红外传感器（例如，苹果手表后盖上的传感器）就可以测心率。因为佩戴在身体上的设备在推挤碰撞时容易松开，所以很多医院为了得到更准确的测量结果，改用红外摄像头监控病人的心跳。Xbox家用电视游戏机利用红外线追踪玩家身体活动的幅度，实时了解他们兴奋或无聊的程度，并据此推出了一个又一个新游戏。

在生物医学层面上，情感更难遁形。验血可以发现与害怕、紧张、疲劳有关的生物化学物质，验汗也可以实现相同的目的。在美国国防部的资助下，通用电气公司成功地研发出Fearbit，它是一种可以吸附到皮肤上的无线传感器，外形与邦迪创可贴相似。朝向皮肤的那一面是纳米结构，可以吸附特定的生化物质。如果这些生化物质的含量升高，它还会发出警报。“嗅探”空气中化合物的传感器的体积非常小，可以安装到手机中。用石墨烯制造的传感器具有非常高的灵敏度，可以检测浓度在10亿分率量级的分子。早前的一项研究表明，我们甚至可以通过人的呼吸检测他的紧张情绪。

在具体环境中综合使用多种情感传感器，可以产生革命性的效果。例如，麻省理工学院媒体实验室情感计算小组的几名研究生提议研发“AutoEmotive”（自动电子功能）系统，将几种既有的传感器嵌入汽车操作系统，改善驾驶员的健康与安全状况。在方向盘上安装传感器，可以监控与紧张情绪有关的重要生物指标，包括掌心出汗、心率、呼吸和手掌抓握力等。利用麦克风监控所有语音的音调和音量，可以判断警

报针对的是暂时性情况还是不断加剧的沮丧情绪。一台车载记录仪可以提供驾驶员微表情的精准数据。如果驾驶员表现得十分紧张，数据服务商就会给他推荐一条更通畅的路线，或者让汽车音响播放舒缓的音乐。驾驶员可以从汽车仪表盘背景灯的颜色变化了解自己的情绪状态，并根据生物反馈做出更明智的决定。AutoEmotive的目标是帮助人们在极易导致“视野狭窄”的高度紧张的情况下做到应对自如，这与埃里克·霍尔维茨为美国国家航空航天局地面控制台设计数据优化显示系统的初衷不谋而合。

在思考如何将情感分析应用到决策活动中时我们必须清楚，关于在特定情感状态下身体内部有何变化的问题，心理学家还没有形成一致意见。分歧最大的问题与情感体验的主观性有关。当前的局面与个人的经历对情感反馈的影响到底有多大？如果表现出害怕的几个特征，比如呼吸与心率加速、流汗、血压升高等，一定是因为害怕吗？出现这些状况，或许是因为你恐惧、震惊，或者感到焦虑不安、心烦气躁，但也有可能是因为你刚吃了一颗糖，而且正在锻炼。

保罗·艾克曼指出，解读情感时须防范“奥赛罗的错误”。在莎士比亚的戏剧《奥赛罗》中，奥赛罗指责妻子苔丝狄蒙娜与卡西奥有染，并告诉她已经派人杀了卡西奥。看到妻子脸上害怕与痛苦的神情，奥赛罗认为这表明她真的有罪。他想，很显然，她感到害怕是因为奸情被揭穿了，她感到痛苦则是因为她在哀悼死去的情人。艾克曼指出，苔丝狄蒙娜在那一刻确实表现出了害怕与痛苦的情绪，但是原因与奥赛罗猜测的并不一样。她感到害怕是因为丈夫妒火中烧、失去理智，她感到悲伤是因为她无法自证清白、自知难逃一死。奥赛罗犯下的令人扼腕的错误说明了一个事实：检测某种情感的生理指标比较容易，而发现其背后的原因却难得多。在利用情感数据进行决策时，无论解读这些数据的是人还是机器，都必须时刻牢记奥赛罗的教训。

面部表情、语音线索生理学数据都是真实的信号，情感识别系统可以从中发现我们大多数人都无法发现的规律。如果可以实时获取经过挖掘的情感数据，我们的生活将会大大改观，但是，风险也会因此增加。你是否想了解自己在第一次约会时或者面试之前、之中和之后的情感状态？检测任一阶段的情感状态，都有可能对接下来的行动产生深远的影响。在面试时，如果面试官告诉你他正在使用情感检测应用程序，你的情感状态是否会发生变化，你会更加紧张还是更加自信？在这种情况

下，你通常会竭力隐藏自己的情感，但如果应用程序利用你脸上的微表情来寻找“蛛丝马迹”，你的所有情感反应肯定会暴露无遗。

我在前文中指出，交流各方都应该有权查看交流记录。如果你打给客服代表的电话被录音，你就有权得到这份录音。但是，由于受情感检测程序监控的交流越来越多，我们无法准确地判断仅仅获取这些原始录音对我们是否公平。如果企业利用语音数据探测你的情绪，并且根据分析结果采取不同的方式处理你的来电，那么它们应该为你提供哪些信息呢？如果你真实的情感体验不同于算法的解读，又会导致什么样的结果呢？

此外，如果我们希望借助情感状态的精炼数据，改进我们与亲朋好友或同事之间的交流，仅凭戴在手腕上的传感器或者对准脸部的摄像头是无法实现这个愿望的。我们还需要想办法充实传感器数据，比如，详细描述并公开分享我们的感受，为机器检测的生理指标添加个性化标签。为了深刻了解我们的行为规律，并帮助我们更好地做出决策，我们可以心甘情愿地公开表露哪些情绪和情感呢？

特克斯勒消逝效应与专注力

通过音频和视频记录，不仅可以探测到你的位置与情感状态，还可以精准地推断出你的关注点。在你凝视某个目标时，用普通摄像机即可轻松地实时了解你视线的大致方向和凝视的时间。转头时，面部识别系统可以辨认出刚刚进入你的视线的人或物。英国兰卡斯特大学与德国马克斯·普朗克研究所的研究人员共同建立的系统，同样利用普通摄像头跟踪人们的视线在广告屏上的运动轨迹，针对他们关注的内容推送商品信息。很多时候，凝视不是一种有意识的行为，而是对周围刺激因素的一种无意识反应。因此，了解人们关注的内容具有非常重要的意义。

摄像头还可以侦测出注意力的微小变化。在为美国国家航空航天局优化信息显示系统的过程中，埃里克·霍尔维茨和他的同事碰到了一个难题——认知负荷。所谓认知负荷，是指处理信息和解决问题时需要付出的注意力。根据丹尼尔·卡尼曼和杰克逊·比提（Jackson Beatty）的研究，瞳孔的相对直径可以反映人在完成任务过程中的认知负荷的大小。在接收新信息（例如，倾听一连串数字）时，瞳孔会放大；在报告这些

信息时，瞳孔会收缩。任务的难度越大，瞳孔的变化越明显。

眼球的微小运动可以揭示很多认知过程。在你“死死地”盯着某个目标时，眼睛仍然需要不停地转动，因为视网膜上只有非常小的一部分（不到视网膜面积的1%）才有辨识细节的锥形感光细胞。此外，大脑也需要处理盲点（视神经穿过视网膜的位置）周围的信息。这些原因产生的眼球运动叫作扫视运动（幅度非常小时则叫作“眼环微小扫视运动”）。扫视的频率为每秒5～100次，幅度只有几度。从扫视的方向、幅度与速度，可以看出注意力的变化情况。此外，从视线停留时间可以得知大脑处理场景中的某些信息所需的时间。视线停留时间是注意力的一个信号。

想要准确率达到100%，眼睛每次最多能看5～7个字母，因此阅读过程中眼球必须完成一系列的扫视运动。我们在阅读时，视线停留在熟悉字词上的时间要少得多。“回视”运动（指眼睛来回扫视已经看过和处理过的信息）表明正在接收的信息令阅读者感到困惑。日本大阪府立大学凯坤泽（Kaikunze）开发的凝视追踪应用程序，可以将注意力分为从“漫不经心地浏览”到“全身心沉浸其中”等若干等级。凯坤泽希望未来的应用程序可以捕捉到让读者感到困惑的字词，并迅速给出解释。此外，应用程序还可以根据人们阅读文本时的投入程度，评估是否“可以打断”其阅读活动。

因为运动幅度非常小（不到1度），所以肉眼难以实时观察到微小扫视运动。不过，包括瑞士Tobii（心拓英启科技公司）在内的几个企业，成功地研发出眼球运动追踪专用设备和软件，捕捉并分析这些微小扫视运动。这些设备通常是用发光二极管（LED）发射的红外线，将设计好的图案投射到实验对象的眼睛上。尽管人眼看不见红外线，但红外摄像头可以监测到视网膜上的反射光，从而推断出眼球所在的位置和运动方向。

Tobii公司的眼球运动追踪眼镜可以应用于“实地研究”，例如，监测人们在商场货架前逗留时会关注哪些商品，或者探寻尚无语言能力的婴幼儿是否（以及如何）拥有感知和认知技能。Tobii公司的另一套系统可以连接电脑显示屏，因此无须使用眼镜。但是，这项技术还需要解决几个难题。比如，其他光源（例如阳光和白炽灯光）可导致红外传感器捕捉到的信号受到干扰。当照明条件发生变化、感知或情感受到干扰时，瞳孔放大的幅度会有所不同，因此凝视追踪与关注度检测的难度也会增

加。

在使用眼球运动追踪设备准确地监测视线停留时间与微小扫视运动之前，必须对设备进行校准。在实验或工作环境中，根据要求，用户可能需要做一次完整的校准练习。不过，研究人员棋高一着，想出了一些让用户无法察觉的校准方法，例如，在屏幕上播放一些肯定会吸引用户眼球的图片。

通过研究，Tobii公司找到了将眼球运动追踪设备收集的数据与人类认知的其他生理线索关联起来的方法。这些生理线索，包括心率、呼吸速率、皮肤电反应（皮肤导电性）、脑电图（EEG，即检测神经系统活动的脑电图）等，在人受到刺激时都会发生变化，对全身心投入和兴趣强烈有明显的预示作用。汇总多种生理传感器数据并加以分析，眼球运动追踪设备就可以根据视线的变化情况，推断出人们的情绪反应。

眼球运动追踪设备收集的数据经过挖掘，可以帮助人们改进注意力训练的效果，提高学习成绩。若干研究表明，经过训练之后，新手可以在富有挑战性的情境中实现专家演示的眼球运动方式。其中一项研究邀请了一些学生在计算机辅导系统的帮助下学习计算机编码，并在他们做习题集时对他们的眼球运动进行了追踪。初学者往往会死盯着一小部分内容，一边演算，一边反复浏览这部分内容；有经验的人在设计编码时，眼睛则会不停扫视，捕捉大量信息。科研人员认为，让初学者学习经验丰富者的凝视模式，可以加快学习进度；还可以使用眼球运动追踪设备，以突出的方式标示出初学者容易忽视的重要信息。

医学上利用类似的眼球运动追踪系统训练专业人员看X线片，检测囊肿和肿瘤。初学者不仅可以向经验丰富者学习，其他初学者的凝视模式对他们也有帮助，因为在看X线片时遗漏的重要信息各不相同。这些眼球运动追踪数据同样可以挖掘。将凝视模式与计算机从X线片中提取的数据相结合，机器学习系统就可以预测出很多人类常犯的错误。最终，在眼球运动追踪系统的帮助下，医生可以做出更准确的诊断。

利用手机和电脑自带的摄像头，可以进行简单的眼球运动追踪。在用户看收集屏幕时，有的智能手机应用可以判断出用户查看的具体位置，并通过可选功能在用户看到屏幕底部时自动向下滚屏。鼠标、触摸屏与语音界面的发明让人机交互情况发生显著变化，在未来的几年里，机器的凝视控制系统的精准程度将不断提高，人机交互必将再次发生翻

天覆地的变化。2015年4月，苹果取得了一项技术专利，即用摄像头和红外传感器检测眨眼的运动，以及眼球运动的时间与方式。利用这些数据，我们可以解决“特克斯勒消逝效应”（Troxler effect）惹出的麻烦——凝视屏幕上某个物体（例如光标）较长时间后，就会发现该物体“消逝不见”了。

图4–2 初学者和经验丰富者在编码时的注意力模式
资料来源：转载自埃米尔·谢雷吉·纳哈尔（Amir Shareghi Najar）、安东尼娅·米特洛维奇（Antonija Mitrovic）与克罗什·内沙迪安（Kourosh Neshatian）发表在《计算与信息技术杂志》上的论文“眼球运动追踪与研究实例”（2015年第2期，23页）。

我们已经知道，正是因为有了这些数据，苹果公司等数据服务商借助凝视控制系统，不仅可以了解电脑或手机屏幕上有哪些内容，还可以深入了解用户到底对哪些内容感兴趣。例如，由于一些神经系统方面的疾病或障碍（包括阿尔茨海默病、孤独症、读写困难、精神分裂症和多发性硬化症）会影响眼球的运动，因此科研人员正在研究如何将眼球运动追踪技术应用到诊断与病人行为的监测之中。

当然，眼球运动追踪设备还可以发现哪些内容没有得到足够的关注。在完成需要持续投入注意力的重复性任务时，随着时间的流逝，视线会频繁地游离应该关注的内容。其直接原因通常是生理或心理疲劳。为了设计出有效的眼球运动追踪系统，在人们意识到有可能错失重要信息之前帮助他们集中注意力，科研人员正在认真研究人们走神时眼球运动是否会表现出某种明显的规律性。走神的频率大概是多少呢？哈佛大学的心理学教授丹尼尔·吉尔伯特（Daniel Gilbert）设计了一个智能手机应用程序，对现实世界中注意力集中的情况进行了抽样调查。根据人们

的自我报告，他们走神的时间占比为20%～40%。

如果在教育、工作和其他领域引进这些应用，允许组织机构更深入、更精细地观察和分析注意力问题，同样会产生十分显著的变化。人们可以利用眼球运动追踪设备，让注意力不集中的人离开任务小组，或者计算员工在完成某些需要高度集中注意力的工作任务后，应该得到什么样的回报。不过，这种做法与公司管理层密切监视员工一举一动的可怕方式并不完全相同。在疲劳可能导致巨大损失的关键岗位，例如长途货车驾驶员或重型机械操作员，眼球运动追踪技术可能有利于提高绩效。此外，吉尔伯特还发现，人走神影响的不只是工作表现和记忆恢复；据绝大多数实验对象的自我报告，在意识到自己走神后，即使当时所想的是一些愉快的事情，他们也会心情低落。知道哪些因素会导致你走神，或者在注意力不集中时能收到警报，这将大大提升你的工作满意度和人生幸福感。眼球运动和注意力的检测能够为个人带来好处，前提是他们可以得到精炼数据，而且这些数据的目的不是控制他们的活动，而是为他们的决策提供帮助。

在传感器被用于收集人际交互中的注意力数据时，情况更是如此。在过去10年里，麻省理工学院的亚历克斯（桑迪）·彭特兰实验室一直在用“社交计量标牌”做实验。标牌上安装了可以检测运动时间与速度的加速计、红外LED，以及用于检测附近标牌发出的LED红外线并推测目标对象身份的红外传感器。他们还利用蓝牙来测量与邻近标牌之间的距离。通过分析这些标牌发出的信号，桑迪和他的同事们绘制出交互图谱（例如，工作场所的员工交互图），并在上面标注出会面的地点、对象与持续时间等信息。社交计量标牌可以记录下详细的信息，例如，开会时谁与谁坐在一起，谁是焦点物，某个人说话时谁会点头表示赞同。此外，标牌还会检测人们在会议中与独自工作时的姿态，了解他们的注意力投入情况与疲劳程度。

这些标牌也可以通过语气、音调、音量、语速、发言与倾听的时间比例、起承转合的规律等语音数据，识别说话者的言行特点。桑迪说，为了让更多的人参与他们的实验，他们没有记录说话的内容。他认为，无须完整的音频记录，也可以发现有价值的交流规律。

他们关注的不是人们说话的内容，而是他们说话的方式。例如，为了使对话的氛围融洽和谐，或者避免冲突，人们常常会不自觉地使用模仿策略，在对话中套用对方的腔调。因此，从交谈时人们迁就某个人说

话模式的情况就可以看出这个人的影响力。说话流畅者（几乎没有“嗯”、“啊”等口头语或停顿，也很少被人打断）常常是那些被奉为专家的人。语音也可以表现出发言者的投入与兴奋的程度，例如，语速加快、音量提升时，就会给人一种富有活力的感觉。

将这些社交计量数据加以汇总，就可以看出团队的凝聚力和个人在这个组织中的地位（地位的高低与名片上的头衔无关）。桑迪认为，这些传感器捕捉到的信号比人们的自我报告和外部观察更加真实可靠。

桑迪指导过的博士研究生、社交计量解决方案公司的联合创始人兼董事长本·瓦贝尔（Ben Waber），曾经与美国银行合作，调查该银行的呼叫中心在人员聘用、经验、培训等方面表现优异的原因。社交计量标牌数据表明，该呼叫中心的几支优秀团队经常举行非正式聚会，而且似乎没有感受到工作的压力。基于这些发现，研究人员建议进行A/B测试，安排呼叫中心的部分团队同步休息，使他们有更多的开展非正式交流的机会。结果，这些团队的工作效率变得更高，绩效提升了约25个百分点。

我们已经知道有无数传感器正在记录我们的感受与我们关注的内容，但到目前为止，还没有传感器可以做到真正理解我们的思想。不过，有的传感器，比如功能性磁共振成像（fMRI）设备，可以观察在决策过程中人脑的活动情况。神经系统科学家对人脑进行时间分辨率约为1秒、空间分辨率约为1毫米的感官刺激，然后利用fMRI技术观察大脑内部的血液流动和氧气输送的情况，就可以确定发生应激反应的部位。接收到相悖于我们的信念的信息之后，大脑中通常负责处理“冷静推理”任务（数学题等不涉及情感的任务）的部位不会产生任何反应。此外，科研人员还可以借助fMRI技术，观察那些应邀执行评估或决策任务的人大脑中的哪些部分变得活跃。他们发现，在很多情况下，决策活动的开始时间要早于他们预计的时间。

fMRI所需的强磁场是由超导磁体提供的，超导要求整个系统的温度降到接近零摄氏度。因此，fMRI设备绝不可能安装到手机上。不过，研究人员正在通过其他办法窥探大脑内部的活动。其中的一个实验性方法是利用小型无线近红外光谱（NIR）传感器，检测大脑皮层的血液流动情况。尽管NIR技术检测血液流动情况的方法不同于fMRI技术，读数也不是非常精确，但是NIR设备的便携性远高于fMRI设备。如果能把fMRI设备的测量值与可以在“户外”工作的NIR设备的测量值结合起来，或许

会带来意想不到的收获。技术创新必将为我们了解大脑内部的工作机制打开一个又一个新窗口。

“奥赛罗的错误”告诉我们，描述人的生理状态与推断在复杂的现实世界中激发这种状态的具体原因，两者完全不是一回事。如果科研人员可以用现场记录校准实验室记录，那么我们必须制定一些指导方针，对发射与传感技术［包括植于皮下的射频识别（RFID）芯片，口袋大小、可以实时完成生物DNA测序的纳米装置，以及目前想象不到的各种各样的传感器］所记录数据的用途进行规范。

一次杜撰出来的“度假之旅”

> 我一直告诫人们要注意当前的行为，因为谁也不知道未来世界将如何分析关于这些行为的数据。
>
> ——布拉德·坦普顿（Brad Templeton）

两年前，荷兰学生齐拉·范·德波恩（Zilla Van de Born）为完成学校布置的作业，设计了一次非常有意思的“度假之旅”，骗亲朋好友说她打算去东南亚各地徒步旅行。在接下来的5周里，她在脸谱网上发布了一系列用Photoshop图像处理软件处理过的照片，分享这个她杜撰出来的探险之旅。从照片看，她今天在尝试潜游（其实是在公寓大楼的游泳池里），明天在品尝当地美味（其实是在附近的饭店里），后天又在参观寺庙（其实是在她的家乡阿姆斯特丹）。她总是在午夜时分上传照片、发表评论，因为地球的另一边这时候是白天。和父母亲用讯佳普打视频电话时，她会把卧室的窗帘全部拉上，并挂上圣诞节灯饰。家人都没有怀疑她，直到她“回家”并把实验的事告诉他们，她的家人才知道真相。德波恩说：“这个实验的目的是告诉大家，我们在社交媒体上公开的信息都经过了筛选、篡改，歪曲现实是一件十分普通而且轻而易举就能做到的事。所有人都知道模特的照片被处理过。我们在生活中也经常篡改事实，但我们常常对此视而不见。”

尽管在网上分享的传感器数据可以筛选、修改，而且的确有人这样做了，但是像德波恩的实验那样创建一个让人深信不疑的替代存在，却并非易事。为了不让亲朋好友生疑，德波恩必须不停地上传假照片，琢

磨与他们视频聊天时讨论的话题。此外，在那5周里，她大部分时间都要躲在公寓里。如果外出（比如，去寺庙拍照片），她就必须不厌其烦地化妆。她别无选择，否则朋友或者脸谱网（或其他数据服务商）的照片识别软件就有可能发现她的位置。但是，大多数人都不可能花这么多的时间和精力掩饰自己的身份。

篡改传感器数据的难度肯定会不断增加。未来，情感识别算法可能会发现德波恩在所有度假照片里都面带“泛美式微笑”，亲朋好友就会因此怀疑她不是真的开心。但是，这并不意味着人们从一开始就会怀疑亲朋好友在欺骗他们，而是我们在同陌生人交流时，可能会不停地描述传感器数据发给我们的真实信号，到最后已经可以信手拈来地使用这些数据了。如果人人都想通过篡改社交数据去欺骗他人，就必然会导致他们利用传感器数据来探查我们到底是谁、身在哪里、过得怎么样。

在本章的开头，我引用了杰夫·格雷因为拍摄奥兰多警察而被捕的例子。很多警察部门都有秘密的车载和随身记录仪视频数据库，但是公众只有通过冗长的司法程序才可以取得使用权。2015年夏天，洛杉矶警察局为警察们配备了7 000个随身摄像头，并且宣布，除了用作呈堂证供以外，不会公开这些摄像头记录的内容。但是，这种不对等性并不是最令人难以接受的。这些随身记录仪不会自动记录警察的对话，何时打开的决定权掌握在警察手中。此外，为数不多的人（警察局的核心人物）拥有自由使用数据库的权利。即使在他们的行为需要接受调查的情况下，他们也可以在写事件报告之前，调出并查看随身记录仪拍摄的所有视频。这样一来，在陈述事件时，他们就可以充分利用视频记录的“盲点”，逃脱纪律处分。在规则面前，警察与他们宣誓保护的对象为什么如此不平等呢？

为了与这种明显的权利不对等现象做斗争，美国公民自由联盟提供了一个云平台，允许每个公民自主上传他的言行的实时记录，以备不时之需。其中一个应用程序可以向在场的其他人发出警告，提醒他们某个事件正在被记录，以免他们用不同的视角记录该事件。

有的地方已经开始接纳透明性这个原则了。2014年12月，西雅图警察局为12名警察配备了随身记录仪。其首席运营官告诉《纽约时报》：“在讨论如何处理这些视频时，有人问道，‘普通公民看我们警察的视频做什么？’”该部门决定不把这些视频收进私密数据库，而是将它们上传到YouTube视频网站上。为了不触犯现行法律，他们还开发了一

种视频编辑算法，对人脸进行模糊化处理，并删除了音频。

尽管这些随身记录仪拍下的视频都经过了编辑，却为人们了解警察的执法行为提供了前所未有的便利。如果愿意，创业型数据侦探可以分析这些信息，从中找出该城市警务工作的规律性，例如警察通常会采取何种方式接触嫌犯、证人和其他公民。这些数据将变成当地居民建设美好家园的一大利器，他们可以据此，建议警察局改变训练或警务程序，并通过A/B测试了解这些措施是否可以改善警民之间的交流。

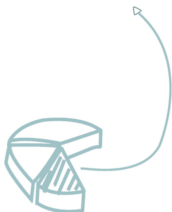
遗憾的是，这些应用程序并不能保证这些数据不会被别有用心的人利用。但是，尽管我们无法阻止人们收集我们的点击鼠标、浏览网页、添加联系人、交谈、走路、眼球运动、喘息和说话等行为的数据，但是我们可以要求拥有与社交数据相关的一系列权利，以确保最大限度地实现数据的透明性和主动性。



第5章
计算隐私效率与数据回报

提高数据挖掘过程的透明性

当你要求查看你自己的数据时，能看到些什么？



> 并非所有能够量化的东西都很重要，并非所有重要的东西都能量化。
>
> ——威廉·布鲁斯·卡梅隆（William Bruce Cameron）

通过前文的论述，我们知道现在我们每天创建与分享大量的数据，也了解到数据服务商的产品与服务对我们的生活产生的影响。我们会不断地创建数据，大多数人也会不断地分享数据。想要数据服务于我们，重点不是对我们的数据实施控制，而是要求数据服务商在控制台上为我们留下一席之地。

我们还探讨了如何提高数据的透明性与用户的主动性。但是，我们大部分人都无法熟练地使用编程语言，不能对整个数据挖掘过程进行核实与破译。我建议用户可以使用一些简单的方法，包括获取呼叫中心的通话记录，将自己脸谱网动态信息中的“头条新闻”与所有朋友的发帖与评论所形成的完整数据流进行比较。

即便采用了上述工具，在没有其他辅助的情况下，你仍然无法对收

集和分析过程中每一个字节的原始数据进行核查。你不得不依靠数据服务商为你解释你的数据的意义。

掌握了有关自己的特征、人脉关系、环境的所有数据之后，无论你是否清楚你自己的愿望是什么，数据服务商都能越来越准确地发现你的愿望。无论机器给出的推荐建议有多好，无论我们对推荐商品有多么喜爱，我们都要保持独立性，包括对推荐建议予以拒绝或调整。这需要我们利用工具了解数据服务商是如何做出推荐建议的，从而实现某种目的，例如修改数据挖掘程序的默认设置与假设。

有些数据服务商注重为个人提供产品与服务，有些数据服务商注重为公司与组织提供产品与服务。这些数据产品和服务将为用户带来更大的透明性与主动性。你所要做的就是选择合适的数据服务商，由他们提供工具帮助你提高数据的透明性与用户的主动性，评估数据服务商回馈给你的好处是否合理。

我们近期才认识到社交数据可能对自己不利。我认为需要制定一整套标准，作为我们对数据挖掘进行评价的标准。以下6项权利为此提供了框架。

提高数据挖掘过程透明性的两项权利：

1. 访问自己数据的权利。

2. 检查数据挖掘过程的权利，包括

a. 查看数据安全审计的权利。

b. 查看隐私权效率评级的权利。

c. 查看数据回报评分的权利。

提高用户主动性的4项权利：

3. 修正数据的权利。

4. 对数据进行模糊处理的权利。

5. 开展数据挖掘实验的权利。

6.自主导入和导出数据的权利

这些权利有助于我们“阅读”数据与了解数据挖掘的特征，“写入”我

们的指令，与数据服务商进行交互。如奥地利裔英国哲学家路德维希·维特根斯坦（Ludwig Wittgenstein）所言："我的语言的局限性意味着我的世界的局限性。"如果我们对某个事物没有概念，我们就会看不见它。对于我们来说，为了理解数据挖掘并用它指导我们的交流，我们就必须学会新的语言。如果我们没有工具对数据服务商的分析与推荐加以评估，我们就不得不凭空想象我们的数据是如何受到不同的解读或使用的。仅访问原始的社交数据意义不大，只有了解如何使用社交数据才能产生效果。

与透明性与主动性相关的权利需要转化成具有可操作性的工具，在你行使这些权利和使用这些工具时，必然会创建更多的数据，数据服务商可能也很愿意分析这些数据。与数据服务商的每一次互动都是一个新的数据点。

我们将在下一章探讨用户的主动性权利。但在本章中，我们需要深入了解用户所需的两种透明性：访问自己数据的权利与检查数据挖掘过程的权利。第一项权利有助于用户查看与解释他自己的个人数据。第二项权利能使数据挖掘过程更加透明，便于用户发现数据服务商处理与使用数据的特点。

用户访问自己数据的权利

无论数据权利的重点是透明性还是主动性，访问数据都是一切数据权利的基础。如我在引言部分所说，个人能够访问他自己的数据，已成为包括美国和欧盟成员国在内的许多国家的标准。但我们目前对"可以访问数据"的范围界定过于狭窄，没有考虑到在性质上已属于社交数据的那些数据，比如与他人的数据交织在一起的你的数据。

想一想你能访问的某个数据范围，以你的金融交易数据与信用记录为例。某些国家的政府已允许个人访问这些数据并修正错误，因为这些数据是金融机构判断人们能否获得信贷的依据。在美国和英国，庞大的消费者信贷资料中心负责收集和分析有关个人负债与信用卡还款行为方面的数据，每年都会向客户提供一份个人信用报告复印件。它们鼓励你检查这些数据，并在发现任何错误时通知该中心。如果你没有申请贷款，却发现许多以你的姓名与住址提交的贷款申请，这就是你的身份被

盗用的迹象。该中心根据你是否按期偿还贷款的记录是迅速偿清短期债务还是累积了高额或循环债务、信用卡账户的开通时间、申请新的贷款的次数，以及你名下的信用卡、贷款与抵押的整体情况，将你的金融行为与财务状况表示为信用评分。它们会告知你，你在去年有哪些行为给你加分或减分，还会通过描述性分析具体说明它们为各种行为赋予的权重。你从中可以了解到，信用评分中有30%来自及时还款，有10%取决于你的信用卡欠款与其他长期性贷款的情况。

如果你收到的信用报告表明由于你经常逾期还款，导致你的信用“风险高于平均水平”，你就需要通过提前或按时还款，努力提高自己的信用评分。你可能认为自己的信用评分是一个数字，即你的金融信用评分（FICO）。但每一个中心给你的信用评分是不同的，因为它们会独立计算你的评分。《纽约时报》统计至少有49个不同版本的信用评分体系，它们依据的不仅是各中心收集的数据和收集数据的方式，还包括你申请的贷款种类。此外，据《财富》杂志称，即便同一家中心对同一个人也不只给出“一个”信用得分，因为“每一家机构”（即考虑是否为你办理新的信用卡、贷款或抵押的银行）都会调整各种参数。这种类似于单向镜的做法，导致你无法像审核数据的银行工作人员那样查看这些数据。

真正的透明性能让你看到不同金融机构对不同范畴的信贷数据所分配的权重，换言之，你能了解到金融机构是如何看待你的信用记录的。通过这种方式访问自己的数据，就能知道申请贷款时应该优先考虑联系哪几家银行。

访问数据的权利还应包括立足于所处环境使用数据。最重要的两个环境是你过去的情况与你的同龄人的情况。对比你过去的情况与你现在的情况，以及你的情况与你的同龄人的情况，就会有所收获。如果没有任何分析、对比、解释工具，单纯访问原始数据是没有意义的。试着对比你今天的行走方式与你过去的行走方式，看看能收获什么。当你行走在办公室里的时候，步态识别软件能记录你的行走方式，还能识别出你随着年龄的增长而发生的步态变化，这可能是慢性背痛或严重肌肉萎缩症的早期征兆。由于尚处在发病初期，你可能注意不到自己的健康出了问题。但我们希望能获得这种早期预警。

你可以将你的数据与他人的数据进行比较，以便从中收获价值。你应该选择谁的数据作为比较对象呢？你要将自己的数据与同在一个科室就诊的其他病人、同一家银行网点的其他客户，还是你所在部门的同事

进行比较吗？如果你把自己的数据与较小群体中的其他人（例如，你工作单位的10多名同事）的数据进行比较时，你可能会依据数据分布情况推测出其他人的相关信息，其他人也能据此推测出你的信息。

实际上，你的许多数据都与其他人的数据交织在一起，也就是说“你的数据”中并不一定只有你一个人的数据。即便数据表面看上去明显是你的数据（例如，从事数据经纪业务的安客诚公司为客户创建的档案），也会与其他人的数据交织在一起。要知道，安客诚公司的用户档案是以家庭而非个人为单位的。广告商的传统营销思维是以家庭为单位，它们解释了这一选择的原因，即许多购买决定都是在家庭这个层面上做出的。但是，并非所有人都希望家人看到可能暴露自己秘密的广告，例如，少女偷尝禁果怀孕后，她的父亲发现她收到了婴儿服装和婴儿床的优惠券，进而发现她怀孕了。

家庭成员共用一个通信地址，而人与人的社交数据的密切程度更甚于此。你会对朋友的脸谱网发帖做出评论，投入时间和精力去创建并分享数据，还会表达你的个人情感与喜好。但无论什么原因，帖子的作者有可能改变主意并决定删帖。那么，你的评论会发生什么改变呢？如果某个人的数据得到其他人的回复，他有权处理这些回复吗？你应当能访问自己参与产生的任何数据，以及查看该数据产生时的背景，但这并不意味着你可以未经他人允许使用他们贡献的数据。

数字化的体验、沟通已成为我们生活的重要组成部分，在我们与他人乃至外部世界进行交互的过程中，我们实际上参与了数据的创建。你在访问数字化“遗产”时应当遵循什么原则？这可不是一个无聊的问题，2015年，脸谱网用户中有100万～1 000万人去世，如果脸谱网用户去世，谁有权接管他们的账户呢？这个问题引起了巨大的争论（和纠纷）。有些账户实际上已应其直系亲属要求被删除。在此种情况下，该账户中所有由多人共同创建的数据（手动标签的照片和对话）均会消失，这是对参与创建这些数据的人的劳动成果的不尊重。

2015年，脸谱网开始为用户提供“遗产契约”功能，让用户指定由谁来行使极其有限的权力——检查并修改逝者的账户，这些人就像遗嘱的执行人一样。他可以选择不同的头像照片，可以创作一篇特别的帖子并将其置顶，还可以接受逝者的亲朋好友发出的好友申请。除此之外，其他的数据访问权限均被禁用。理由很明显，访问相当于能使用数据，专门针对逝者的个性化推送对其他人实际上没有太大的用处。如果该账户

产生过多的新数据，账户就不再具有纪念意义，反而会凸显账户“管理员”的性格特征。

考虑到传感器数据，对“你的数据”下定义变得更加困难。在我们的社会中，在公共场所拍照的人将进入相机取景框的人拍下来，根本不必征求对方的同意。拍摄者可能不知道照片中每个人的身份，但脸谱网的DeepFace人脸识别软件却能做到这一点。该软件利用了脸谱网巨大的照片数据库，这些照片已由脸谱网用户或其好友标记出照片中人物的身份。脸谱网的算法极为复杂，能够自动标记出新上传照片中的人的身份，但它只会对与被标记者互为好友的用户显示这种“计算机标签”。

有些国家的政府认为，这种标签触犯了人的隐私权。实际上，在欧盟对计算机标签提出反对后，脸谱网在欧洲国家主动关闭了这项服务。但为什么允许人工标签，却禁止计算机标签呢？被标记身份的人可以将这两种标签从搜索结果中删除，人工标签与计算机标签都会给人带来帮助或伤害。

无论是谁发的照片，脸谱网都会努力标记出其中每个人的身份吗？它必须这样做。原因在于，如果未识别出照片中的人，它就无法及早将某些照片从其算法中过滤掉。难题出现了，如果人们担心照片标签会给他们带来不利，应该怎样解决这个问题呢？政府可以禁止使用计算机标签，但这种一刀切的法规会减少人们从数据挖掘中获得的益处。更好的选择是让人们了解计算机标签的风险与回报。

如果你不认识的人在脸谱网上贴出一张活动照片，你身在其中，还有其他人。人脸识别软件会识别出其中一个人是你，当你的脸谱网好友浏览这张照片时，就会向他们推送含有你的信息的计算机标签。但是，如果无人注意这一标签，你就永远不会知道这张照片的存在。

如果你要求脸谱网提供所有对你进行标签的照片，无论是人工标签还是计算机标签，无论贴出这张照片的用户是否允许你查看这张照片，脸谱网都应向你提供这些照片。由于计算机标签有一定的错误率，因此你会看到被脸谱网误识别为你，但并非你的照片。如果你希望查看每一张对你进行标签的照片，你可能需要对大量的照片进行筛选。为了提高可操作性，你需要按照近似度对这些照片进行排序，并修改那些标签错误的照片。满足上述需要的工具尤其重要，因为通过综合运用大量的照片、视频及其他数据，可以推测出你在不同时间和地点的行动，就像

Vigilant公司通过一系列摄像头定位跟踪汽车全天的行驶路线一样。在行使访问你自己数据的权利时，无论数据源自何处，你都应该能够访问与你有关的所有数据。

你有权看到出现在照片或视频中的其他人的脸吗？如果其他人不是你的好友，你有权看对他们的计算机标签吗？毕竟你们都参加了这项活动，通过询问活动的主办方、摄影者或其他人（也许还可使用谷歌的反向图片搜索工具），你可能会了解到这些人是谁。即便软件对照片中的其他人都进行了模糊处理（类似于西雅图警察局对在YouTube视频网站上播放的探索视频的处理方式），也能识别出他们的身份。但是，提高发现陌生人身份的成本也有好处，它可以防止我们的数据给自己带来不利。假设出现了一种极端情况：某人参加了一项活动，他在房间里到处闲逛并尽量使自己被照相机拍到。之后，他得知所有有他在其中的照片都已被数据服务商识别出来，他可能会利用这些照片中或附加的各种公共信息，发现当时谁在场，其中也包括他想寻找的人，这个人可能是他的潜在客户（会受到骚扰），也可能是犯罪对象（会受到伤害）。计算机对照片进行标记之后，会让人们设定规则：谁可以看到这些标签。这不仅改变了获取这一信息的成本，还限定了标签的使用范围。一般来说，每个人都应选择是否允许他人看到自己身份的标签，还要选择允许谁看到这些标签。

有两个问题导致你访问你自己数据的权利变得复杂，即数据归谁"所有"。"数据所有权"的实际含义是什么？从历史上看，数据的产生者与数据的相关者可能存在利益冲突，例如上述照片标签的例子。但从更基础的层面来说，当下所有权的概念已发生了变化。如果我买了一个苹果，它就是我的，我想怎么处理它都可以。我可以把它切块，也可以把它吃掉，还可以送人或卖掉。但无论什么时候，它的所有权只能归一人所有。当然，你只要咬一口，就不能改变它的所有权了。与此相反，1个字节的数据可以由多人同时使用，而且他们不会把它消灭掉。数据可以同时归多人所有。实际上，数据的所有权并不是指在决定数据的命运方面具有唯一的决策权，例如购买、出售、捐赠或销毁，而是访问数据的权利和使用它们的权利。

考虑到这些，你就会发现为什么获得原始数据还远远不够。我认为，访问你自己数据的权利包括：通过慎重考虑决定能够向他人展示的数据和不能向他人展示的数据，设计出复杂的算法、用户界面与软件，

以及能隐藏你的部分身份信息并提供有细微差别的选择，决定向谁分享数据和分享多长时间。其中的部分内容并不容易实现，包括计算机代码和社交礼仪。但这并不能阻止我们对此的需求，因为我们对自己数据的查看与使用均依赖于此。

用户检查数据挖掘过程的权利

我们如何确定数据服务商对我们数据的挖掘，能给我们带来不错的回报且风险不大。为实现全面的透明性，你不仅需要拥有访问自己数据的权利，还应该拥有检查数据挖掘过程的权利。

我认为，你应有权检查数据服务商的社交数据生态系统是否诚信、健康，根据数据服务商抵御安全攻击的能力衡量其“卫生”程度；测量它对数据的使用效率或删除隐私所需的时间；评估你能从分享你自己的数据中所获得的回报大小。在医疗领域中，我们明白即便保持良好的卫生状况也无法预防百病。这三项措施同样如此，其目的旨在使数据挖掘过程变得更加透明。

以餐馆卫生评分为例，许多地方都要求对餐馆进行卫生检查，依据明确的二元评判体系。如果评分达到最低要求就可以继续营业，否则就会被勒令停业。实际上，公共卫生检查人员在对餐馆进行卫生评分时，通常会将得分分成从“差”到“优”的几个等级。

通过检查评分制度（而不仅仅是合格或不合格），有关部门就能够对截然不同的两项目标（实现公共健康与经济活力）加以权衡，以满足社会大众的需要。如果某人的免疫系统较弱，他就会选择去卫生评分最高的餐馆就餐。即便这个人只是一个特例，有关部门仍需要考虑卫生条件对他产生的负面影响。同时，有些人可能愿意接受稍差的卫生条件（比如卫生评分为良），以换取更低廉的价格或更美味的菜肴。

卫生检查无法预防罕见但危险性极高的风险。即便人们拥有极其强大的免疫系统，食用了伤寒带菌者烹制的食物后也会患病。常规卫生检查将确保餐馆工作人员经过培训，工作时戴手套，生病期间不上岗。

所有这些卫生评分的概念都适用于数据产生与分享过程中的取舍。无论我们做出什么决定，都存在一些消极和积极的因素，会产生意料之

中与意料之外的结果，包括使用哪些数据帮助我们做出决策。意料之外的结果是你没有预料到的，甚至是你想象不到的。它十分罕见或几乎不可能发生，如在你就餐的饭店中出现了伤寒带菌者。意料之中的结果是你能够控制和预料的，比如在就餐后必须埋单。无论是意料之中还是意料之外的结果，都可能会产生积极或消极的影响。伤寒带菌者与餐厅的账单都产生了消极影响，只不过生一场重病的风险比就餐费用更难以计算。

在数据的世界中，最重要的意料之外的消极结果或风险之一就是安全受到威胁，即访问你数据的人打算滥用数据，或者可能对你造成财务或其他方面的伤害。这些罕见的事只要发生一起，就可能会同时影响数百万人。由于你的信息被数据服务商反复加工和处理，会导致你的隐私权一步一步地被侵犯，这是常见的意料之中的消极结果。由于其他人反复询问数据服务商，使某些数据与你产生关联的概率增加，不可避免地会或多或少侵犯你的隐私权。但是，你有多少隐私被使用或“侵犯”，取决于数据服务商的政策与规程。这可以通过隐私效率评级来确定。

通过牺牲一定的隐私权换取你意料之中的积极结果或收益，这可以通过数据回报评分来评估。数据服务商的差异在于它们对你的原始数据的使用效果、搜索结果排序和你的偏好的匹配程度。对数据服务商的比较应依据三项标准：你的数据面临的安全风险，隐私效率评级与数据回报评分，你可以据此对数据服务商进行选择。

	消极结果	积极结果
意料之中	隐私效率评级	数据回报评分
意料之外	数据安全审计	意外收获

你也许注意到了针对数据挖掘所产生的意料之外的积极结果，我并未提出任何建议和具体措施。我在表中将这种情况称为“意外收获”。意外收获是指你通过使用社交数据，在约会地点遇到令自己怦然心动的伴侣，在领英网上成功获得梦寐以求的职位，对未确诊的病情发现了症结所在，或是找到了人生的重大问题的答案。尽管我肯定数据服务商将在这些改变我们人生的重要决策方面提供越来越多的帮助，通过个性化推荐与排序产生幸福、惊喜的结果（这是你人生中值得见证的时刻），但它们过于主观化、个人化，无法表示为某个数字。

现在，让我们探讨用于评估数据服务商的三项标准。

数据安全审计

似乎每隔几个月，就会看到大规模数据泄露的新闻。每一项科技都有其优势和劣势。大多数科技在带来积极的改变的同时，也会带来风险。比如，开车可能会车毁人亡；将数据分享给数据服务商，也可能使你面临安全风险或遭遇黑客攻击。

马克·古德曼（Marc Goodman）是《未来犯罪》的作者，也是联合国、北约组织、国际刑警组织的顾问，他强调安全漏洞不应当被视为发生概率极低的事件。15%~20%的世界国内生产总值（GDP）涉及有组织的犯罪活动、贩毒、人口贩卖与卖淫、窃取交易数据和侵害知识产权，互联网为这些犯罪活动提供了巨大的机会。

最著名的数据泄漏事件包括：零售商塔吉特百货与易贝公司的交易数据遭窃，摩根大通的财务数据遭窃，索尼电影娱乐公司的职员受雇数据遭窃，医疗保险公司Anthem的患者数据遭窃，菲律宾选举委员会的投票数据遭窃。它们都上了新闻头条并引发了人们深切的焦虑。但是，事件发生后，很少开展明确的公开审计以找出到底哪里出了问题，也没有公开讨论怎样提高公司或组织数据的安全性。数据是在从一个集线器传输到另一个集线器的过程中遭窃的吗？能从机构内部找到薄弱环节吗？在向某个数据服务商分享自己的数据时，如何判断这样做是否“足够安全”？

在数据泄漏事件发生后，公司经常表示自己在高科技的安全攻击面前无能为力。有时，这种辩解是合理的。例如，在索尼电影娱乐公司受到黑客攻击后，其发言人称：“任何关于本公司事先应做好抵御此次攻击的准备的说法都有失偏颇，并且忽视了联邦调查局的重要发现与看法。”联邦调查局网络处处长对参议院提供的证词称：“攻击者使用的恶意软件……可能绕过了私营部门目前所采用的90%的网络安全防护，不客气地说，政府的网络安全防护也如此。”黑客使用的恶意软件相当于伤寒带菌者：公司有理由也有能力在数字化安全防护方面进行巨大投入，但依然易受黑客攻击，这属于意料之外的、消极影响较大的结果。

对于收集与分析社交数据的每一家公司而言，数据安全审计必不可少，而且结果应当提供给用户。但是，大多数数据服务商都没有积极性分享检查结果，因为它们担心一旦公布安全审计结果，就会更容易受到

黑客攻击。如果公布的安全评级较低，就相当于在自家门前竖起一块广告牌，上面写着“房子大门没上锁”，这纯属开门揖盗之举，更不用说明确公布数据服务商在安全防护方面的薄弱环节了。但犯罪分子通常会根据目标的价值选择攻击目标，而不是它们的安全薄弱环节。

索尼电影娱乐公司聘请了一位著名的网络安全专家凯文·曼迪亚（Kevin Mandia）保护公司的资产安全。许多公司为了保护敏感信息（包括价值数百万美元的好莱坞电影发行数据和数百万条客户的信用卡卡号数据）的安全，都聘用了安全顾问。公司的信息安全与用户的信息安全是一致的，它们既不希望重要数据遭窃，也不希望因受到安全攻击而遭受巨大损失。但目前，大多数用户都无法判断自己存储在某个公司的数据受攻击的可能性大小。他们无法对各个公司的数据安全程度进行比较。这就是我们要求数据服务商开展数据安全审计，并发布其结果的原因。

作为数据安全的组成部分，数据服务商需要与用户沟通他们采用的安全标准。美国电子前线基金会指出了超文本传输协议“内在不安全性”的根源，就是以非加密方式传输数据。遗憾的是，这依然是大部分网站的默认方式。数据服务商应当采用安全超文本传输协议，它在客户与服务器之间建立了加密链接，使外部攻击者难以拦截通信过程。

安全审计将检查数据服务商的雇员对数据的访问情况。数据服务商会要求两次或两次以上的授权，即雇员除了输入密码之外，还需输入手机应用程序或专用设备提供的临时验证码，这比不要求进行此项验证的数据服务商的数据安全性更强。此外，对每一次访问的记录和分析也有助于提高公司的安全评级。此类记录不仅能够帮助调查人员寻找异常迹象并追查原因，而且能被鼓励雇员遵守公司规章、勤奋工作。众所周知，如果人们知道自己的行为正在被记录，他们就会改变自己的行为。

但是，许多安全漏洞并非因为软件存在缺陷，而是公司内部人员存在人性的弱点——雇员会心怀不满、怨恨，未获得充分培训，或因为过于忙碌而对工作敷衍了事。用户应当对数据服务商的数据安全等级拥有更大的知情权，包括要求数据服务商证明负责处理数据的雇员值得信任。金融机构需要对所有应聘者进行背景调查，包括他们是否在之前的工作单位有过欺诈或舞弊行为。数据服务商还必须评估其雇员可能造成的安全风险。开发人员可以接触到大量用户的原始数据，公司可能要求他们编写或编辑代码，却很少逐行检查他们的代码。如果数据服务商发

现某个开发人员疏于遵守安全规程或提交的代码漏洞百出且对此不作为时，就必须严厉惩罚他。

实际上，某些最具危害性的安全漏洞大都是因为不严谨的数据处理活动而非恶意攻击所致。如果某个政府部门负责管理的硬盘中储存了7000万名美国老兵的医疗与军队退役记录。在硬盘发生故障后，工作人员会将硬盘交给生产商并要求对其进行修理。如果无法修复，生产商就会将硬盘交给第三方做回收处理。然而此时，数据仍存储在硬盘中。负责管理这些记录的机构会做出声明，在其与硬盘生产商签订的隐私条款中，后者已承诺会对此数据保密，但犯罪分子并不关心这些法律细节。这种数据处理的方式和负责管理数据的机构所做出的声明都令人无法接受。

在对任何数据泄露风险进行定量分析的过程中，都应严格评估雇员对数据处理安全操作规程的了解，就像在餐馆卫生检查的过程中需检查员工对食品安全规则的了解程度一样。这取决于公司的文化，而非员工个人。

数据也有可能被机器人窃取，这些机器人经常潜伏在面向个人用户的网站中，以便从中窃取信息。应对的方法是建立相应的安全系统，探测出异常活跃的账户并将其关闭。例如，如果某个用户每天24个小时、每周7天不间断地在线并一刻不停地翻阅网页时，这肯定不是人类用户。这些机器人及其幕后的操控者变得越来越狡猾，现在它们会采用不容易被发现的信息窃取方式。在这场没有硝烟的战斗中，如果数据服务商训练机器学会识别未经授权的数据提取行为时，就可获得更高的安全评级。

即便数据服务商实施了精密的监控，机器人仍然能够窃取数据。在线患者社区网站PatientsLikeMe的用户有过沉痛的教训。这家网站是为人们管理医疗情况的社交网站，用户在这家网站的论坛中分享的数据可能十分敏感，比如，他们的临床诊断结果、目前的健康状况、使用处方药与非处方药的情况、药品进行副作用、病情预测等。许多用户就因患有慢性疾病所付出的生理与心理代价进行交流，例如多发性硬化症、艾滋病、创伤后应激障碍、抑郁症等。通过相互分享信息，能够产生神奇的效果——用户了解到他人如何处理类似的问题，借鉴他人的经验教训，并将他人的治疗情况作为衡量自己的治疗情况的标尺。

虽然有些用户在网站上使用了假名，但也有些用户使用了真实姓名，还有些用户在个人档案或签名栏中公开了自己的电子邮箱地址和其他身份信息。虽然这更便于其他用户直接联系他们，但也使别有用心的人很容易猜出他们的真实姓名与身份。因此，当PatientsLikeMe宣布网站论坛受到机器人的侵入时，你可以想象用户的震惊程度。机器人偷偷地从该网站论坛窃取数据，并提供给尼尔森调查公司，再由后者用于为某家制药公司或医疗设备公司所做的市场分析。PatientsLikeMe关闭了这些机器人账户，但是大约有5%的发帖已被复制。定期进行的数据安全审计应当评估公司的数据处理安全规程，以预测其受到机器人账户窃取数据的可能性。

在安全漏洞受到攻击之前就应该努力寻找并弥补安全漏洞，上述这些举措应成为其中的组成部分。例如，自2011年起，任何人只要发现程序漏洞或安全隐患并将其报告给脸谱网，就可以获得一笔奖金。脸谱网通过这种方式已找到了2 000多个安全漏洞，并向发现这些漏洞的黑客支付了400多万美元的奖金，但这只占脸谱网安全预算总额的很小一部分。脸谱网通过对安全漏洞的风险程度进行评估，以及公司是否已意识到这一问题的存在，来决定支付给每个“正义”的黑客的奖金。目前，最大的一笔奖金是33 500美元，奖金得主是一名巴西人，他能够通过用户忘记密码后重置密码的请求代码中的漏洞侵入脸谱网的服务器。并非所有的数据公司都有资源实施此类安全举措。但是，实行奖金计划可以高效地消除安全漏洞。

如果你思考一下“物联网”，就会认识到黑客入侵的后果可能极为可怕。飞机、火车、汽车上的计算机系统对海量数据进行分析，从家到医院、嵌入式计算机网络无处不在，这导致人们面临生理风险。在第4章中，我们说过GPS干扰仪可能将人们引至错误的地点。在一次令人大开眼界的实验中，得克萨斯大学工程学教授托德·汉弗莱斯利用GPS干扰仪和遥控装置，在船员毫无察觉的情况下控制了游艇的导航装置。在另一次实验中，两名黑客证明他们可以通过车载娱乐系统控制一辆切诺基吉普车的“转向、刹车和传动系统”。

这些实验正在改变公司、机构和组织对数据安全的看法。2015年，梅奥医院聘用了12名“正义”的黑客并组成团队，看他们能否入侵控制数百家医院的医疗急救设备的网络。其结果令人震惊、发人深省。这些设备的易攻击程度令人难以想象，它们的安全防护措施形同虚设，虽然预

定时间为1个星期，但黑客（其中很多人是网络安全防护领域的知名人士）无须多长时间即发现了所有的安全漏洞。有些设备使用的仍是原厂设置的默认密码，只要有人有兴趣入侵，不费吹灰之力即可做到。在阅读了该团队的报告后，梅奥医院修改了其采购政策与规程，要求所有医疗设备都必须遵守严格的安全协议。但是，这份安全协议在医疗领域中还远达不到标准协议的水平，通常仅根据费效比分析即做出数据安全监控决定。在索尼电影娱乐公司遭遇著名的黑客攻击事件之前，该公司的首席执行官曾于2007年称不愿“投入1 000万美元去预防可能发生的100万美元的损失”，这反映了他的惯性思维。此次数据泄漏最终造成了惨重的损失，仅“调查与补救费用”就高达1 500万美元，而且严重损伤了索尼公司的商业信誉。有一个方法可以减少安全漏洞检查的费用，即建立独立的行业组织，对全行业开展数据安全审计，并以某种方式对黑客进行认证。同时，如果用户坚持要求采用较高的数据安全标准，不报名参加此组织的公司的安全防护成本也会上升。

通过对著名的数据泄漏事件进行梳理，我总结出5个实现数据安全的措施，要求数据服务商做到。第一，要像卫生检查一样，必须达到最低要求，否则不能对外营业。例如，必须安装最新的安全软件，并为所有已知的安全漏洞打上“补丁”。第二，数据安全不仅限于安全代码，它还与人有关，因此要建立尊重用户及其数据的公司文化。外部组织进行的数据安全审计可以检查数据服务商的总体健康状况，并对其安全规程与措施进行评级，其中包括公司员工是否接受过数据处理安全方面的培训。第三，邀请正义的黑客定期入侵数据服务商的网络和计算机，检查是否存在不为人知的安全漏洞。在理想情况下，审计机构还要评估数据服务商对异常情况的反应时间，并提出具体的改进建议，这有利于数据服务商及其用户。第四，必须对所有的数据服务商采取统一的审计标准，以客观评估其数据安全情况。第五，在某些类型的数据被泄漏时，应当有能力对其带来的潜在危害进行评估。如果你在驾驶一辆计算机联网的轿车时被黑客攻击，其造成的危害可能比你的信用卡发生欺诈交易的后果更严重。

当审计方检查数据安全清单、检查并测试数据服务商的安全规程时，它们会给合理的举措加分。用户有权查看服务商最近的安全评分，还要能查看之前的历次得分，以便将最新的得分放在大背景下比较和考量。如果某个数据服务商的审计结果有所改善，就表明它对安全防护措施进行了投入，或其员工的安全培训收到了效果。如果得分下降，就可

能表明它最近出现了安全漏洞或疏于培训新聘员工。

即便数据安全评级或风险评级的结果很糟糕，如果因数据服务商的疏漏而导致发生数据泄漏，数据服务商也不能逃避任何法律或道德上的责任，由此产生的损失必须由数据公司与用户共同承担。否则，在发现竞争对手的安全评级较低（或虽然安全评级较高，但产品与服务很糟糕）时，公司可能认为没有必要提升自身的安全性，尤其是未来对个人用户造成的伤害可能无法确定为由某一次数据泄漏或其他安全问题所致。针对此类问题，如果公司不愿主动赔偿，政府或法院可以对公司进行罚款并用于补偿用户。

由于数据服务商已越来越广泛地深入我们的生活，有关数据安全的更多信息需要提供给公众，这一点很关键，而且要在数据泄漏事件发生之前。实际上，由于社交数据越来越多地用于识别人们的身份、声望、心理状态，数据泄漏的风险将威胁到每一个人。

隐私效率评级

信息泄漏是毁灭性、意料之外的事件，但用户使用数据服务商的产品或服务也要承担定期的、意料之中的成本，包括隐私权逐步受到侵犯。众所周知，你需要向数据服务商提供个人数据，才能从数据服务商那里获得数据产品与服务。隐私权是一种资源，在用用户数据创造产出的过程中被数据服务商消费。

与各种资源一样，隐私消费的效率可能各不相同。微软研究院的辛西娅·德沃克称，在分享自己的数据时，必须对隐私权的损失情况进行量化。辛西娅认为，这种“差分隐私”的目的在于建立数据系统，并确保该系统中的所有用户都不会在分享数据的过程中得到直接的负面结果。辛西娅将此归结为两个问题：“对于同等程度的隐私权损失，哪一种技术能提供更高的精度？对于同等精度的技术，哪一种技术能提供更好的隐私权保护？”

用户在分享自己数据的过程中获得了回报，但数据服务商必须对用户的隐私权损失实施管理。这有利于将数据服务商视为一种生态系统，通过关注整个生态系统的健康而并非关注该系统中每个个体的健康，对

该系统进行最佳维护。精度与隐私权之间的取舍并非针对某一个人，而是每一个人。你在选择数据服务商的过程中，就应该了解它们通常会迅速还是缓慢，低效还是高效地消费隐私权。

隐私消费的速度与效率类似于工程学与环境科学中的“燃烧率”概念。工程师制造出一个烧木材的炉子，它可能会快速燃烧大量木头却无法让房间暖和起来。它虽然是个能正常工作的炉子，但它的效率不高。把更多的木头投到炉中也许能继续给房间供暖，但远达不到理想的水平。数据可能不再是稀缺资源，但隐私权是稀缺资源，而且会越来越稀缺。隐私权就像木头一样，常常还未产生多大的价值就已消耗殆尽。

现代化的木材燃烧炉应获得认证，它的整体燃烧效率需满足一定的要求——60%~80%，并根据理论上的最高燃烧效率（100%）进行改进。隐私效率评级也可以按照类似的方法进行。100%的隐私效率是指在理想状态下达到一定程度的精度时，隐私权的损失程度最低，而且使用的数据为必不可少的数据。例如，导航服务必须获得人们当前的位置和目的地位置，才能为他们指路。

在寻找改善其产品与服务的方法的过程中，数据服务商针对应收集哪些用户数据不断做出决策。亚马逊公司根据用户的点击与购买记录建立了数据库，并利用该数据库向客户推荐商品，且不需要保存每个客户的个人身份记录。亚马逊关注的是客户从一件产品跳转到下一件产品的轨迹，而不是奥马哈的维罗妮卡点击浏览了这件产品后，又点击浏览了另一件产品。如果人们查看这些数据库，不会发现个人身份信息，这就降低了客户数据遭窃的概率。

我在为交友网站Fridae工作时，分析了用户对其他用户添加的数千条备注，比如，“给我发了5条消息，需要回复”“遇到他了，但不是我喜欢的类型”“化学专业的荣誉毕业生”“看上去比29岁大多了”。这些备注只对添加备注的用户可见，其他所有用户都看不见。我对此类信息十分感兴趣，想知道从这些备注中可以采集哪些类型的信息，以及是否可以将这些信息纳入网站的设计。分析表明，这些备注可帮助用户记住他给哪些人发过信息或与其约会后无功而返，从而避免再次与这些人接触。但是，我们在分析备注的内容之前已将所有的用户名删除了。因此，在改进该网站服务的过程中，我们降低了隐私的消费量。我们团队中的任何人都无须知道某个用户的具体偏好，只要知道其备注的模式，并思考如何在网站中增加新的功能即可。

如果不考虑效率因素，就很容易打造出一款功力强的机器。如果是为一级方程式大赛的参赛车辆设计引擎，结果可能耗油量巨大。汽车厂商曾连续几十年都不太关心汽车的耗油量，因为汽油价格低廉，似乎取之不竭；汽车买主则对其他问题更感兴趣，例如车辆的外观、性能和价格。20世纪70年代的石油危机大大改变了汽车厂商在做引擎设计时所考虑的因素，更注重性价比，政府也强制规定汽车的燃油效率必须更高，客户也更在意耗油量。

由于对引擎的要求不同，燃油效率（在美国按每加仑汽油行驶的英里数计算）可能相距甚远。在起步、停车频繁且车速较低的城市里行驶，与在高速公路上的匀速、高速行驶相比，前者的燃油效率会低很多。天气因素和其他油耗因素（例如空调）也产生了影响。美国环境保护署通过在实验室中测试5种驾驶情况，将其各总结为一项燃油效率评级。每一款车型都包括在同一份清单中。美国环境保护署称："修建了平整的测试场地并在实验室条件下对各种车型进行测试，以确保测试结果前后一致、准确、可重复、公平。"

令人遗憾的是，虚假的燃油效率评级可能会让人信以为真。诚然，各种燃油效率评级中都包括人们可能不会遇到的驾驶情况。你住在酒店里，如果发现房间温度过高，就会调节房间温度，但温度似乎仍没有变化。这时，你会请工作人员来检查。当工作人员检修完毕后，温度显示屏的温度读数下降。如果你觉得房间依然温度过高，你可能会怀疑工作人员"修好"的是温度显示屏，却并未修好温度调节装置。这时，你可能会要求工作人员提供温度计，测量房间温度。

但是，在很多情况下人们并没有计算能力或测量能力，无法对测试结果的可信度提出质疑。机器的构造极为复杂，可通过对它们进行配置让其在某些环境中更高效地工作，其中也包括在这些机器接受检查的环境下。美国环境保护署在实验室中对车辆一氧化氮的排放情况进行测试，而且这种环境可以被车辆传感器感知，于是大众汽车公司的工程师借鉴了这一点。由于他们无法在提高大众汽车公司的各种柴油引擎的燃油效率的同时，使一氧化氮不超过排放标准，于是开发出一款软件以降低车辆在接受测试时的一氧化氮排放量。直到西弗吉尼亚大学的研究人员在实际道路上测试大众汽车的一氧化氮排放情况，才揭穿了这场骗局。

数据挖掘的复杂程度不亚于汽车。即便我们清楚地了解到必须提供

哪些数据方可获得某些产品和服务，我们也难以判断隐私效率评级的可靠性。但是，发布数据服务商的隐私消费情况就像发布汽车燃油效率一样轻松。无须对每加仑汽油行驶英里数进行计算，我们的目标是计算出数据服务商用每单位的隐私损失所解决的问题数量。

在实践中，需要用一组测试对隐私效率进行评估，这类似于美国环境保护署的一组标准化燃油效率测试。这组测试将对数据服务商进行剖析，了解其识别某个人平均需要的互动次数。互动次数越多（在隐私损失发生之前进行了多次询问），表明数据服务商的隐私消费效率越高。

许多人正在研究保护隐私权的同时保持数据用途的方法，例如辛西娅·德沃克与英国实业家约翰·泰森姆（John Taysom）。他们的工作表明，有可能开发出一种工具测量数据服务商对用户隐私的使用情况。约翰·泰森姆已为一些有趣的发明申请了专利，它们旨在减少因提供数据产品与服务所必须消耗的隐私量。泰森姆还认为，我们既不能依靠数据服务商的自查自纠，也不应依赖政府机构的监管。他说："人的寿命预计能延长到100年，至少在发达国家如此。但公司的寿命相对而言要短得多。政府在保护个人数据方面没有进行完善的记录，这种情况需要改变。政府也没有正确的数据管理结构，可能需要记录100年甚至几代人的数据，例如基因数据。"针对数据安全性，应该由数据专家组成独立的组织，为我们评估并公布隐私效率情况。

我们在了解与管理因获取数据产品与服务所损失的个人隐私方面，尚处于起步阶段。未来可能会出现一些惊人的创新，例如，环境科学家在研究气候变化时，会观察地球上碳资源的燃烧速率。燃烧速率不仅要通过每年消耗的碳量进行计算，还要测算出在不造成地球生态系统失衡的前提下每年可以消耗的碳量。为鼓励人们减少对碳的使用量，某些国家给其境内的公司规定了年度碳排放量额度。如果该公司的实际碳排放量低于其排放指标时，即可将未使用的碳排放量额度出售给碳排放量大的公司。如果某家公司的碳排放量过大且无法买到碳排放量额度时，就得支付罚款。这提高了公司的生产成本，迫使其减少碳的使用量或提供比竞争对手好得多的产品，让买方甘愿支付更多的费用和环保成本。组织与个人也可以主动为环保活动捐款，以此抵消自己的碳足迹。

今后，我们可能会允许数据服务商对隐私使用额度进行交易。但在此之前，我们必须先开发出能客观衡量它们所消耗的隐私量并将结果公开发布的工具，该工具还要能评估我们对各种隐私消耗率的满意程度。

数据回报评分

你在提供自己的数据时，如何估算所获得的收益呢？从概念上说，数据回报和隐私消耗率都能反映数据服务商对数据的使用效率。隐私消耗率衡量的是预期成本，即你在使用服务数据商的服务时你的真实身份的暴露程度。数据回报衡量的是预期收益，即相对于所分享的数据所获得的价值多少。它们都能帮助你判断你所获得的信息产品与服务，是否值得你将自己的数据提供给数据服务商。

许多数据服务商会要求我们提供过多的信息，而我们无法判断这些信息是否有必要提供。就像你在初次约会时，对方提出了20个问题让你回答，但却不告诉你关于他的任何信息。这种约会一定不会有美好的结果。但是，许多数据公司刚与你建立联系时，就采用了这种做法。用户必须在提供数据给数据服务商之前，通过某种方法评估此举的潜在收益。数据回报评分为我们提供了工具，它可以衡量数据服务商对我们的回报大小。

关于提供数据所获得的收益，每个人的认识都很主观。对某些人来说，在脸谱网上分享自己孩子的一张照片属于重要的信息发布，但在某些人看来，这没什么大不了的。有人认为认识某个朋友（例如，后者分享了他欣赏巴赫的无伴奏大提琴组曲后的感受）非常有价值，但也有人认为这种交往纯属浪费时间和精力。虽然只在人们将数据提供给数据服务商并体验了其产品与服务后，才会真正了解到他们的数据所带来的回报，但人们应当能根据某个数据服务商过去和当下的用户所获得的平均回报来判断是否使用这家数据服务商。数据回报评分是用人们获得的平均回报（获得的信息产品与服务的价值）除以平均投入（提供的个人数据）计算出来的。

那么，怎样测量提供数据所获得的回报呢？我们可以先查看某人提供的数据是否对数据服务商的产出有所贡献。如果没有，这些数据的回报即为零。但在大多数情况下，数据都会产生一定的作用，而且测量数据回报是比较棘手的问题。

让我们看一看如何计算出分母，即用户为数据服务商提供的数据量。这笔投入所产生的回报通常以美元或美分为计算单位：你在某个项目、投资组合或公司投入一美元后，所产生的回报是多少？正如我们在

第1章中讨论的，将一串数据分享给数据服务商，很难对其进行合理的定价。数据的投入不能直接用金钱表示，也不能按照其字节数计算回报。

但是，可以通过用户的活动或关注度来衡量确定他们对数据服务商所做的数据投入。测量用户的关注度比测量用户为数据服务商投入的时间更复杂。例如，浏览器中某个页面被打开，这并不能表明用户确实在浏览这一页面，除非他做出了某些举动：点击鼠标、移动鼠标、检索、评论、上传及下载（或者该用户允许数据服务商通过其电脑的内置摄像头对其进行监控，并将视频流发送给数据服务商，以便数据服务商对用户进行直接观察）。某些数据服务商要求新用户在使用数据服务前必须填写某些个人资料。如果这些数据是必不可少的（例如，只有填写地址才能收到所购商品），就可以通过它获得一定的回报。但某些数据被采集的原因仅仅是营销人员认为自己需要客户的人口统计信息，这对用户毫无价值。

测量数据投入需要考虑的另一个因素是查看数据的创建与分享方式。专门为数据服务商创建数据，例如填写调查问卷或上传用户头像，其工作量比分享点击鼠标、移动鼠标与检索等行为的视频。在计算某个用户的数据投入时，显性数据应当比隐性数据的权重更大，这是常见的做法。

这种权重分配机制还表明，当用户尝试使用另一家数据服务商的服务时，通过脸谱网或以类似的方式登录该网站或应用，可减少其因获取个人化的推荐建议所进行的投入。脸谱网还向某个应用分享了相关数据，例如可将用户的头像照片分享到优步或来福车的约车平台，以便驾驶员与乘客能够互相辨认出对方；脸谱网还可以分享你的好友收藏的歌曲，以便音乐流媒体服务商Spotify能够将这些歌曲加入你的播放列表，这样你就不必浪费时间创建他人已创建的数据，也节省了你的精力。

空中食宿网提供的是房屋合租和客房合租服务，它的网站或手机应用要求用户关联其社交网站账户。实际上，空中食宿网已告知用户，它需要访问大量的用户个人数据，包括政府发放的身份证、网上身份证、头像照片、电子邮箱地址、电话号码，以验证用户的身份并帮助用户彼此间建立信任感。对于该网站而言，在线身份包括用户的社交网络，因为在脸谱网上伪造出数百名相互验证的朋友关系要比伪造个人档案困难得多。在这个案例中，我们可以发现，为了通过房主或房客的身份审

批，需要较多的社交图片信息，但并不需要提交个人的所有数据。你账户中的其他数据可用于改善空中食宿网的服务，例如，你喜欢宠物吗？你喜欢每天早上出门锻炼吗？空中食宿网可以通过分析你在脸谱网上与朋友的互动，为你找到合适的房主或房客。但让房客和房主明确表明自己在选择房客或房子时关注的重点是什么，这样做效率更高。

现在，让我们来讨论分子，即用户从数据服务商那里获得的价值。该价值可能表现为改善沟通、改善速配、改善决策等形式，但难以对改善程度进行量化。用户的收益无法通过他们花在网站上的时间进行衡量，更不要说对快乐感和满意度的衡量了。我们需要采用另一种方法，那就是统计网站每个月的活跃用户数量，即在过去30天内有多少人访问了该网站或应用。但是，访客数量可能只能反映该公司最新的营销活动的效果。

更有效的方法是针对数据服务商定期统计和分析的用户参与数据，考察其中的多个因素，即使用的时近性、使用的频率、使用的种类。用户上一次访问该网站或应用是多久以前？用户平均多久访问一次该网站或应用？用户在该网站或应用中参与了多少项不同的活动？时近性取决于数据服务商提供的服务类型。假设用户平均每6个小时访问一次谷歌网站，你平均每6分钟使用一次谷歌搜索引擎，这并不意味着你从谷歌公司获得了更多的好处，而表明与你的其他活动（包括你的睡眠活动）相比，你的搜索活动过于频繁。只有将谷歌和用户使用的同类搜索网站进行比较，时近性才能成为有效的指标。但是，如果某人昨天使用了相亲应用软件，该应用就会给他靠前的排名。与一个月甚至一年前使用该应用的用户相比，这名用户可能获得更靠前的排名。更多地参与能够提供更大的收益。使用频率指某人每天、每周或每月使用某家数据服务商产品或服务的次数，也可以将不同时间的使用频率进行对比。如果人们目前对某个数据服务商的平均访问频率比一年前更低，就表明他们的数据回报更少。最后，使用种类表明数据服务商依据其收集与分析的数据为用户提供的产品与服务的范围。在理想情况下，数据回报评分可以让用户查看使用性时近性、频率、种类等数据，还能查看不同数据服务商给各个因素分配不同权重将如何改变最终得分。

有了分子与分母之后，就能计算出每个人所获得的数据回报。对所有用户的数据回报进行汇总，在不考虑每个人对该数据服务商的使用频率的情况下，即可得出每个人的平均数据回报（如果我们先将所有用户

获得的收益相加，再除以所有用户的投入，那么活跃用户将比不活跃用户获得更大的权重）。如果数据服务商的数据回报评分小于1，就表明用户从该数据服务商处所获得的收益通常低于对其的投入。这听上去不是等价交换，更不是一笔理想的交易。但数据回报不仅取决于这两个变量，还要考虑数据的决策价值。

对用户行为进行观察时，应该用有关用户动机的定性数据对数据回报分数加以补充。例如，可以要求用户总结自己使用某个数据服务商的经历，这更像一项客户满意度调查，目的是对产品或服务的收益以及客户继续购买或使用它的可能性进行量化，这有助于了解高频率访问的实际情况。数据服务商知道用户多久会再次访问它的网站或使用它的应用，而且可依据点击和查询情况推测出用户的目的，但无法准确地知道原因。通过简单的调查问题，可以了解更多的深层次信息。例如，“我们注意到你今天在三个不同地方访问了本公司的网站。你多次访问是因为你找到了自己需要的产品或服务却被中途打断，还是因为你虽然没有找到自己需要的产品或服务，但认为我们推荐的商品非常好呢？”

净推荐值法可能对数据服务商极具吸引力，这确保数据服务商的利益与用户利益的一致性。这种方法要求个人对自己推荐某个公司的可能性打分，每项分值为0~10分，总分范围从负100分（受访的每个人都是“贬低者”，他们不可能推荐这家公司）至100分（受访的每个人都是“推荐者”，他们会推荐这家公司）。如果调查人员询问某个用户：“你向朋友或同事推荐本公司/产品/服务的可能性有多大”时，自然就会生成一个选项，该用户可进行推荐并分享给社交网络中的好友。在此情况下，用户的利益与数据服务商的利益可能是一致的：反馈意见可以转化为数据服务商用户基数的增长。

用户在使用过程中会逐渐了解到自己能获得的回报，这不仅要与自己的投入相比较，还可以放在更大的范围之内，将其与使用数据服务商所产生的安全风险与隐私成本相比较。用户有权查看并比较各家数据服务商在意外风险、预计成本、预计收益方面的估值，据此决定选择哪一家数据服务商。

用自己的数据投票

人们在驾驶车辆时，希望一眼就能了解车辆的关键性能指标，例如通过引擎故障灯、机油压力灯、油量计、速度计等，这样就能将主要精力放在开车上，同时还可用余光观察周围的环境，从而安全抵达目的地。汽车仪表盘装置会筛选出最重要的信息并显示出来，以便使驾驶员迅速看到并做出决定。

为了真正实现透明性，我认为我们需要采用类似的标准化仪表盘，以便对反映数据服务商“健康与卫生”状况的三项指标进行检查。新用户在创建账户之前，就应该看到这些检查指标。因此，这些指标需要一目了然地出现在数据服务商的网站首页上或应用软件的说明中。对于老用户来说，这些指标可以与用户的个人页面集成，例如与谷歌现有的仪表板集成。我希望看到“数据服务商的服务商”出现，它类似于导航网站，可以收集数据计算指标，并将结果发布给用户。这需要明确、直观的显示，例如将检查结果绘制成类似于光谱的饼图，并依次标示出绿色、琥珀色、红色部分，表现最出色的数据服务商的评级为绿色，表现最糟糕的则为红色。这也需要成立独立的外部组织，对数据服务商进行检查和评级。

我希望人们在决定是否将自己的数据分享给新的数据服务商时，要看看仪表盘的习惯。仪表盘有助于人们更好地评估自己使用各个数据服务商的感受，并决定是继续使用这家数据服务商，还是尝试使用新的数据服务商的服务。人们对某个数据服务商用得越多，就会越关心隐私消耗率，还会将当下使用的数据服务商与其竞争对手做比较。他还应当使用自动接收通知功能，当这家数据服务商的三项指标中有一项低于一定水平，或该数据服务商进入“红色区域”时，就能收到通知。

随着我们对数据与数据挖掘的理解逐渐深入，这些指标的具体计算公式也将有所改变。与法律法规一样，最重要的是指导原则；当我们了解到如何最好地利用数据并保护我们的隐私权时，相应的细节也会发生变化。但如果我们坐等数据服务商为我们制造出仪表盘，我们将什么也得不到。我们必须要求数据服务商提供工具，以便对其绩效中与我们有关的各个方面做出评估。

一位用户的数据对数据服务商的经济价值很小，与此相似，一位用户对透明性的要求也很少会引起数据服务商的注意。但是，社交数据革命涉及许许多多。100万乃至10亿用户对透明性的要求不容忽视了，而且现在他们的要求可能更多。因为这10亿人不仅要收发电子邮件，还要

使用各种神奇的工具（这些工具多由数据服务商开发），以便发现、沟通、组织人与信息。通过使用这些工具，我们可以发现希望提高数据处理水平、隐私效率、数据回报的其他用户。

总之，我们可以对不达标的数据服务商施压。我们可以用自己的数据投票，选择能够满足用户需要并提供透明性数据的数据服务商，避开缺乏透明性、擅自将我们的数据提供给他人使用，或给用户的回报极少的数据服务商。

如果这样做收效甚微，我们还可以发起虚拟抵制活动，即大家团结在一起拒绝将数据分享给表现糟糕的数据服务商，除非我们看到它们的绩效出现明显改观。如果数据服务商对我们的要求不予理睬，我们就对政府施压，出台法律法规要求这些数据服务商定期接受检查并公布结果。

目前，我们完全可以无限制地查看仪表盘，但除非我们能根据显示信息采取某些行动，否则我们对数据服务商给我们的回报只能进行有限的掌控。为确保我们能够依据这些信息采取行动，有4项主动性权利至关重要，分别是修正数据的权利、对数据进行模糊处理的权利、利用数据开展实验的权利，自主导入和导出数据的权利。

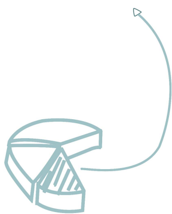


第6章
让数据为你服务

提高用户对数据的掌控力

在数据公司处理你的数据时，

你应该对它提出哪些要求？



启蒙所需要的仅仅是自由，并且我们在这里所讨论的自由，也是所有形式中最不具危害性的，即能够在一切事务上公开地运用理性的自由。

——伊曼努尔·康德（Immanuel Kant）

数据服务商是一台机器，它的运转依赖于人类下达的某些指令。因此，虽然某家数据公司的内部机制完全透明，但也不能确保有关你的数据和你分享的数据都能为你所用。负责设计这部机器的人员可能会告诉你（甚至他们自己也会相信），他们最清楚如何为用户设置各种参数。但你怎么知道，数据服务商会不会像弹球机一样把用户当作球，由数据服务商的负责人根据自己的意愿任意投掷、碰撞、旋转？每当球滚入、击中一条广告或其他付费内容时，这些负责人就能获得授权或奖金，你也能知道这台机器的设计目的是提高命中广告的概率。

这就是为什么数据服务商的透明性不足的原因；我们还必须获得主

动性，能够自由地决定数据服务商怎样使用我们的数据，在数据服务商的控制台上拥有一席之地。

这也适用于我们与数据服务商的互动。对于一些常见的问题，我们可以轻松地将其交给电脑。例如，将收到的电子邮件标记为“垃圾邮件”或“非垃圾邮件”，没有人想像过去那样，电子邮箱的收件夹中充斥着低价出售伟哥等广告邮件。然而，你也会不止一次地发现，你一心等待的邮件被放进了垃圾邮件文件夹，或者更令人气恼的是，某个人收到了你发给其他人的邮件。垃圾邮件过滤系统让你可以选择将某封电子邮件标记为非垃圾邮件，并将其移入收件箱，同时调整甄别垃圾电子邮件的规则。上述反馈均可改进你的邮箱系统的效率。

垃圾邮件过滤系统会通过设置调整，在一个极端（被误标记为“垃圾邮件”的邮件数量很多）与另一个极端（被误标记为“非垃圾邮件”的邮件数量过多）之间获得平衡。在大多数情况下，垃圾邮件过滤系统会根据发件人的数据和进入服务器的所有邮件的元数据，对你收到的每封邮件进行垃圾邮件评分，以便机器能够进行学习。为了改进系统，你的电子邮件服务商也应该向你提供具体的分析过程，即为何将某封邮件放入垃圾邮件文件夹。此外，你可以选择深入了解垃圾邮件过滤规则，并对其进行调整，以便更好地反映你自己的偏好与沟通模式。

一方面，如果你不想花过多时间手动删除垃圾邮件，而且不担心邮件被误认为垃圾邮件后丢失，你就可以对垃圾邮件过滤规则设置严格的参数条件。另一方面，如果你不想丢失任何邮件，也不介意在检查垃圾邮件方面花费时间，你就可以将参数条件设置得更加宽松。如果你有许多朋友和家人在尼日利亚，你可能不会介意该国垃圾邮件比例高达90%的情况。为什么不给人们更多的自由给机器提供反馈意见呢？这可以使用户在自己的邮件归类过程中拥有更大的发言权。为实现这一目的，邮件服务商需要公布垃圾邮件过滤系统的运行细节并允许用户调整相应的参数，对自己邮件的处理方式拥有一定的掌控权。

提高用户的主动性需要将对数据和数据服务商工作过程的掌控权转移到用户手中。服务商从4个方面赋予用户掌控力：修正数据的权利，对数据进行模糊处理的权利，运用数据开展实验的权利，自主导入和导出数据的权利。修正数据的权利为用户提供了表达自我的机会，对数据进行模糊处理的权利则通过用户的自决权来实现。利用数据开展实验的权利通过给予用户探索的自由提高其主动性，自主导入和导出数据的权

利则给予了用户来去的自由。根据这4项权利开发相应的工具，将改进数据服务商的信息产品与服务；同时，建立后隐私经济，让数据为你服务。

拥有修正数据的权利

人类历史上最早的文字记录大约出现在6 000年之前，当时苏美尔人发明了楔形文字。作为统治阶级的祭司（国王）承担起巨大的责任，他们负责制作、烘干、储存和保管刻有文字的黏土板，其中包括著名的苏美尔泥板文献。这些泥板记载了各种归属权——因税收、租赁、缴费、借贷或贸易而负债的人及其债务，以及对这些归属权和交易活动具有约束力的法律。由于泥板的内容十分重要，因此出现了伪造品，针对泥板中“不可更改”的内容还发生了争斗。苏美尔人决定将这些泥板封存，委托当地的神庙保管。这意味着由祭司来掌控这些数据，决定谁可以获取这些官方记录，谁不可以。所有人都希望祭司能准确记录信息，不会因利益驱使而滥用他们保管的信息。不幸的是，祭司（国王）与精英分子并不总是正确和值得信任的。

皮特·沃登是图像识别初创公司Jetpac的合伙创始人，他有力地证明，今天的人类已经进入了狂热地保护数据的新阶段。我们发现了问题（不准确的数据可能造成的危害），并认为解决问题的办法是确保所有的数据准确无误。但由于如今的数据已达海量程度，不可能保护每一条数据不被修改。苏美尔人曾力求为约100万人实现这一目标，但当他们将信息的掌控权交给一小撮人后，却发现数据仍有可能遭到破坏。人类不可能核查每一个字节信息的真实性。民主德国曾力求为1 600万人口实现这一目标，他们通过斯塔西招募了全国劳动人口中1%的人从事此项工作，结果发现仅凭手工验证远远无法完成此项工作。但是，今天我们可以通过机器学习核查数据的真实性。

我们需要努力实现只保管“准确的数据”这一目标，并帮助用户更好地在记录中留下自己的痕迹。修正数据的权利指主动绑定数据，创造和分享与现有数据相关的新数据，能像其他数据一样进行汇总和分析。人们可采用与个性化商品与服务推送相同的算法机制，这有助于在特定情况下推送个性化很强的信息，包括与其他数据绑定的数据。鉴于目前的数据已达海量程度，计算机能分析大量的、可能相互矛盾的数据，现代

化通信的成本也较低，我们再也不需要将世界划分为非对即错的二元世界。修正数据的权利采用了概率论的世界观。

修正数据的权利有助于产生比欧盟的被遗忘权更大的用户主动性。以私募股权投资人格雷格·林代（Greg Lindae）为例，在欧盟的这项指令生效之后，他立即针对1998年《华尔街日报》中的一篇关于他参加密宗工作坊的文章，要求在谷歌搜索结果中隐去这篇文章中对他的报道内容。这篇文章是一条重大新闻，林代的申请也十分吸引眼球，于是报纸编辑决定跟踪报道林代，并针对他的申请写了一篇报道。具有讽刺意味的是，点击这篇新报道必然会提高“密宗”或“被遗忘权”的搜索排名，而其中都指名道姓地提到了林代。（只在欧盟境内使用计算机检索林代的姓名时，谷歌公司才会隐去他本人要求删除的链接。）林代承认，被遗忘权不太可能成为全球标准，对他来说，更重要的是能对以前有关他的报道发表某些评论。他对《华尔街日报》说：“如果能够增加一点儿背景……就没有问题了，也会更好一些。”绝大部分人无法像《华尔街日报》或为我们“修正”数据的部门那样修改数据。无论我们的数据是否具有吸引力，我们都需要拥有修正数据的权利。

在信息处于公共利益和个人隐私之间的地带时，被遗忘权指令并没有提供明确的标准。因此，各组织面对此类申请时可能会“重写”人们认为对其决策有用甚至是关键性信息，但这却挑战了其他人的“知情权”。此外，到目前为止确定此类申请有效性的唯一方法似乎是采用人工方式逐个评估每一项申请。这就像回到了苏美尔人的时代，由祭司决定保留或销毁哪些泥板。

人们根据欧盟被遗忘权所提出的申请，大多都涉及其他人发布的信息。这是因为如果人们决定不再将本人之前创建和分享的数据公布于众，大多数平台都会允许用户删除这些数据。但与访问数据的权利一样，很多涉及某个用户的数据都是由别人创建的，更准确地说数据是“各方共同所有的”。无论是关于政治还是产品，如果某人删除了对话中的某一部分，就会导致断章取义。

如果修正数据能够使人们从中获益，他们修正数据的积极性就会很高。如果某栋乡间别墅的价值评估出错，别墅的主人必然比政府工作人员更有积极性去修正错误，但这仅是在某些情况下。如果房主认为评估人员高估了房价，她可能希望对此价格信息进行修正以降低房产税。如果她打算将房屋出售，她可能乐于见到高估的价格，因为这可以抬高房

屋的售价。相较之下，评估人员没有太大的积极性去检查评估结果，因为这不仅耗费时间，也许还要承认错误。只有在房价被低估，并且由政府付费委派评估人员再次评估房价时，才有可能修正原先的记录。

如果数据可能会对你产生危害，修正数据的权利就变得尤为重要。你手机的定位数据可能表明你在某个时间待在新泽西，而你当时实际身处曼哈顿，因为你的手机连接的是哈得孙河对岸的基站。此时，你可能需要证明你当时确实在曼哈顿。你可能会找出其他数据（例如录像）证明你确实身在纽约，并将这段录像与当时的位置进行绑定，同其他人表明这是相互矛盾的数据。未来，数据将会超出我们的掌控范围。例如，安装在公共场所的摄像头，可以帮助评估你是否符合申请职位或贷款的资格。你无法阻止这些数据的产生，但你有要求将个人数据与之绑定的权利。

此外，无论关于你的数据是什么，你都应当有权进行修正，例如反驳、解释或否认，对查看此数据的所有人高亮显示，而且你的修正在优先性方面高于其他人。此外，如果某个数据中既涉及你也涉及他人，你就应当通过某种方式表明修正这个数据对你的重要性，帮助数据服务商确定如何对你的修正与其他用户的修正进行排序。如果这种修正服务不收取任何费用，社交数据平台可能会被此类申请搞得精疲力竭。可通过推出虚拟收费服务来解决这个问题，即每个用户都可获得一定的积分，用于各种修正申请。

还可以依据验证情况确定修正的权重，包括验证数据来源和其他用户的反馈。正如第2章所讨论的，红迪网通过识别在投票系统中作弊的账户名以保证投票的公正性，例如通过对相似的IP（网络之间互连的协议）地址进行验证，识别来自于同一个用户或同一个群体的投票。好评或差评也属于数据修正行为。但是，互联网上的恶意攻击可能会伤害，甚至摧毁某个人、某个在线社区的声誉。为了维持健康的生态系统，一种方法是要求用户在每一次申请修正数据时都要提供个人身份，既可以是真实姓名，也可以是常用的假名。但长期使用的身份并不能保证修正数据的真实性，就像在法庭开庭前，无论证人在保证“本人证词均为事实”时看上去多么真诚，也不能保证其证词的真实性。如果将修正数据申请与某个人绑定在一起，以便进行问责并减少负面因素时，这在内部举报的情况下又无法奏效，因为透露内部证人的真实身份会导致他们有性命之忧。最近的例子包括维基解密证人对政府秘密的曝光、巴拿马避税

文件的曝光及Labor Link系统对工厂的工作环境的曝光。

修正数据行为的元数据可用于验证，例如创建修正数据的时间与地点。视频与音频记录中的背景噪声也能透露出电流的频率。在中国和美国，交流电（AC）的标准频率是50赫兹，而在欧洲和日本，标准频率则为60赫兹。在这两种情况下，当电网负荷发生变化时，频率的变化存在细微的差别，这种差别足以透露出具体的时间和地点信息，而且可以精确到分钟。在美国与加拿大有4个主要电网，每个电网在针对负荷需求量做出反应时都有各自独特的“频率签名”。通过将某段记录中背景噪声的频率波动与当地电网的频率波动特征进行比较，就可能识别出录像拍摄的具体日期与时间，以及大致位置。

这个例子有点儿像区块链的概念，记录中包含了创建数据的时间与地点信息，并且无法删除。区块链概念是为虚拟货币比特币所开发的数字化分账系统。区块链是过去所有交互与交易的永久性记录，它涉及数据中嵌入的某些数据，这一点很重要。因此，它始终在接收数据，而且无法拆解、修改或删除。区块链记录了每个比特币所有者的数据，即便许多比特币用户选择用假名绑定“钱包”，它也依然能确保某枚比特币不会同时发生多次交易。该系统的建立旨在通过去中心化、分布式过程公开记录每一笔交易，任何人都可以阅读数据并写入数据。这一设计的基础在于，只要某个字节的数据被分享，就再也无法删除，因为这个数据会出现多个副本，它们被保存在整个互联网的多台机器上。每一次转账和数据修正都是透明的，并可以追溯其发生时间。此外，人们还可以在区块链中嵌入备注，以标注某次交易的背景。

区块链也可以仅在群体或组织内使用，不对外公开，某个群体或组织（被称为“联盟”）可以通过它阅读（写入）所有的历史数据。这对于医疗数据可能是极具吸引力的选择，因为只有患者、医生、获得批准的家人允许修正记录。完全公开的区块链更加透明、可信，因为交易数据可在整个网络中得到验证。如果有人想要篡改任何重要数据，就会被其他人发现。与此相反，联盟可以更快地处理交易，因为在验证和保存交易数据的过程中涉及的参与者更少，但他们也更容易勾结在一起，数据遭到篡改的风险更大。无论是公开还是私营的区块链，都保存了所有的交易历史，达到了前所未有的明确问责程度，这在有人用你的数据给你带来风险时尤为重要。区块链可以持续地确保修正行为与产生修正数据的人绑定，具有强力胶效应。

在上一章的讨论中，我们提到数据服务商在对用户创建的数据进行计算时，对显性数据分配的权重大于隐性数据，因为用户对前者投入了更多的精力。修正数据属于显性数据，它可以绑定任何显性或隐性数据。通过修正隐性数据（例如，注明照片上的元数据不太准确），你就能在数据回报计算的过程中提高它的权重，并表明此类修正活动与你的关联度或对你的吸引力更大。

最后，为了让用户拥有修正数据的权利，数据服务商必须投入资源去支持用户修正数据的权利。如果能将用户的注意力吸引到可产生利润的领域中（例如付费广告），可能就会对数据服务商产生较大吸引力。应当将虚拟现实资产用于显示修正数据情况并建立推送它们的系统，这是回报用户的必要的初步措施。

拥有对数据进行模糊处理的权利

对数据进行模糊处理的权利，是指用户有权决定自己所分享数据的详细程度。用户提供的数据越模糊，他从数据服务商那里获得服务的个性化程度就越低。尽管如此，人们仍然有权提出自己的条件，并在不同时间、不同地点选择适合自己的各种个性化程度。如今，我们可以极为精确地测算出人们的各种特征，例如通过全球定位系统或信号灯上的摄像头了解人们所处的位置。但这并不表明我们想要或需要向数据服务商提供如此高精度的数据，实际上，我们应当有权决定自己所分享数据的详细程度。如果我们只能在分享高精度的数据或无法获取数据服务商的服务之间进行选择时，那么这只适用于数据比较粗糙的情况，但现在这种选择再也不存在了。

有时我们需要或想要获得高精度的地理位置信息，有时则不需要。如果你希望及时收到某件商品时，你会毫不犹豫地分享自己的精确位置。如果你拒绝将自己的详细地址提供给达美乐比萨的外卖服务人员，你就无法收到你想吃的比萨。但在许多情况下，我们只需分享精度一般的信息，即可获得我们所需的产品或服务，而且价格更加实惠。你可以要求出租车司机在你目的地附近的十字路口停车，无须向他提供准确的目的地地址，由此产生的成本只不过是多走几分钟的路。在大多数城市，你都可以在谷歌地图中将终点设为你的目的地附近的街道，这样做同样能获得较为准确的路线规划指导。

我们既可以通过上述方法人工更改或模糊处理数据，也可以利用科技手段实现这一目的，对智能设备或应用提供的高精度数据，减少其精确的位数或特征后处理再将其提供给数据服务商。根据微软研究院的埃里克·霍尔维茨提出的模型，人们无论身处何地，都可以对手机传输给数据服务商的地理位置数据设置空间精度，精度的可选范围从1米一直到全球范围内。埃里克认为，精度设置取决于你的具体情况。当你在停车场里寻找自己的座驾或在商店里寻找某件商品时，你可能希望获得最高精度的数据。如果你上班时间却在购物商城中闲逛，你可能不希望透露自己的精确位置。算法可以根据各种变量了解你的偏好，例如当时的时间或地理位置。如果你应邀与一名新客户去咖啡店喝咖啡，你可能希望知道精确的位置。此外，精度并不一定是以物理距离为衡量单位的。如果你身处人烟稀少的地区，就不能将自己的位置模糊为几英里范围内，而应将它模糊为距离最近的1 000部手机的范围内，这样你就不容易被找到。数据素养要求你懂得所提供数据的合适精度水平，以便从数据服务商那里获取你想要的产品或服务。

地理位置不仅仅是调高或调低数据精度。无论是人与人之间的关系、人们在点击与移动鼠标的过程中表露出的偏好与情感、与某种情况的关联程度，还是将某个位置视为私人场所或公共场所，所有这些因素都比拨动开关按钮更复杂。生活并非非黑即白，数据同样如此。

许多个人特征（包括年龄、体重、身高、种族、宗教信仰、雇主、行业与职业信息等）都可以进行模糊处理。例如，你在领英网上浏览其他用户的档案时，领英网允许你对他人能看到你的身份信息的详细程度进行模糊处理，而且系统可以让你预览模糊处理后的个人信息情况，但你在查看访问你的个人档案的用户档案时，也只能看到同等详细程度的信息。如果你是正在求职的女性或少数民族，你可能希望对自己的身份信息进行模糊处理，即只向筛选简历的招聘人员显示自己名字的首字母。经济学家发现，名字看上去像“少数民族”（或外国人）的应聘者，获得面试邀请的概率要低于名字看上去像“白人”（或美国人）的应聘者。

拥有模糊处理数据的权利，还有助于人们在商业背景下更好地掌控自己的数据。在购买商品时，零售商需要了解你所购买产品的具体SKU（库存保有单位）。某件商品的详细特征会透露你作为消费者的许多信息，针对涉及个人敏感信息的购买行为，你可能会要求商家将具体

的商品模糊处理为它所在的商品类别，例如“按摩工具”或“保健品”，也许还可将其归入“健康与个人护理”或“健康与居家”类别中。通过将具体的商品模糊处理为产品类别，可以保护消费者在账户遭遇黑客入侵或出现意外关闭的情况时（忘记退出账户）发生尴尬。当然，对产品数据的模糊处理会影响你收到的商品推荐信息，因为你的购物记录与具体商品之间的联系已不存在。但是，在这种情况下，获得更少的个性化推荐信息可能恰恰如你所愿。

为了测量对数据进行模糊处理的权利的实际效果，你必须先创建数据，有时甚至是十分精确的数据。如果民用全球定位系统不够精确，你就无法获得有关行车路线的有效指导。如果用手机打电话，就不得不连接基站，手机运营商显然会知道你的位置。只在你决定将自己的数据分享给一个数据服务商，并对该数据的使用设定某些限制条件时，你才能降低数据的精度。

在某些情况下，你在创建数据时即可改变数据的精度。但是，在“源头”对数据进行的模糊处理是不可逆的，这可能导致你今后无法使用一些数据产品与服务，包括无法获取你想要或需要的各种提示。如果你对自己身份中的某些信息进行模糊处理，你可能就无法进行电子支付。

你可能已决定对数据进行模糊处理，但之后却发现高精度数据与你需要做出的决策息息相关。假设你居住在毒品交易猖獗的地区，你决定将这个地址模糊处理为周边几英里的范围，以免受这个声名狼藉的街区的牵连。但之后，你可能想知道自己居住的环境是否有较高致癌风险。如果该地区的某些建筑物检测出高于正常水平的铅或其他致癌物质含量，但因为你对此地址进行了模糊处理，数据服务商就无法对你的致癌风险进行准确评估。

选择对数据进行模糊处理经常会产生各种结果，你并不总能提前预测到这些结果。亚马逊的金读电子阅读器（Kindle）记录了人们阅读活动的起始页码与结束页码，以及每读一页所花的时间。即便这些数据有助于老师对让学生感到困惑的课程内容进行个性化设计，学生也很有可能不希望老师看到这些数据，这取决于这些数据对他的成绩的影响程度。假设你决定将高精度的阅读数据分享给亚马逊或其他图书推荐网站，以获得个性化的图书推荐，但之后却有可能发现美国联邦调查局的特工出现在你家门前，因为你曾花大量时间阅读的一本书讲述的是波士顿马拉松炸弹袭击案犯是如何制造高压锅炸弹的。这种设想与实际情况

十分接近。

在各种情况下对各个层面的信息进行模糊处理后，对我们乃至我们今后的决定所产生的影响还需要花时间去观察。我们在建立数据模糊处理的心理模型的过程中发现，如果提高数据的边际精度，使数据服务商的产品与服务的边际价值显著提高，用户就可以获得巨大的回报。如果用户能够自主地采用他人已经完美设定的数据模糊处理参数，以此作为自己的首选参数并无视服务商的默认设置，也能从中获得极大的回报。针对手机或计算机的默认设置，用户也许应该将其置换为电子前线基金会、美国公民自由联盟或类似组织推荐的数据模糊处理参数设置。组织可以针对不同类型的用户提供一些数据模糊处理的建议参数设置，并说明其有利与不利的方面。在研究了不同的参数设置后，你就会知道哪种参数设置最适合自己，并对参数设置进行微调和个性化处理。

当数据涉及个人敏感信息时，如何创造出鼓励人们公开数据并“对此负责”（对数据进行更多的修正）的环境呢？例如，如果某人知道自己的意见与老板相悖，可能就不敢实名发表政治评论。

20世纪60年代，加拿大经济学家斯坦利·华纳（Stanley L. Warner）在为自己的研究收集实地统计数据时，就遇到了这种问题。他发现人们经常出于某种原因拒不提供自己的信息，无论你怎么苦口婆心地告诉他分享信息对社会的益处或个人从中获得的回报，都无法说服他们。如果你问他们敏感的问题，例如“你吸食大麻吗”或“你的艾滋病检查结果是阳性吗”，就没有办法知道有多少人在回答时说谎了（因为无法强制进行血液检查）。

华纳猜测某些人会说谎，但他不知道哪些群体说谎的可能性更大。如果调查对象住在附近，很可能会在回答时撒谎，这导致华纳的实验结果产生了无法弥补的偏差。为了在调查对象及其回答之间设立一道防线，他提议对数据进行模糊处理。

他的具体方法是，人们在回答问题之前先抛一枚硬币，如果硬币的正面朝上，他们就要如实回答问题（“是”或“否”）；如果硬币反面朝上，他就要给出虚假的肯定回答。只有调查对象本人知道他是如实回答还是听从掷硬币的结果。当他之后面对自己虚假的肯定回答时，他完全有理由辩称这是掷硬币的结果。没有人会因为他的回答去找他的麻烦，因为他们无法知道他是否在说谎。华纳的方法的好处在于，它在保护调

查对象的同时，也为研究者提供了他们所需的数据。

模糊处理的方法可以应用于人们的原始数据输入和数据服务商的数据输出。凭借对各个要点进行模糊处理却依然能获得数据汇总与分析的回报，可以帮助用户提供建议。

拥有用数据开展实验的权利

数据服务商不断地对自己的设计、设置、排名的算法开展实验。如上文所述，它们还对自己的用户开展实验。如果数据服务商能对我们开展实验，我们就能对它们开展实验。

修正数据的权利允许用户自由地表达意见，对数据进行模糊处理的权利允许用户自行做出决定，利用数据开展实验的权利是探索的权利，它允许用户对种种可能性进行探索。数据服务商的关键作用之一，就是决定产品与服务对用户的显示顺序。数据服务商可依据时近性等参数对搜索结果进行排序，此时时间最近的结果排在第一位；也可按地理距离排序，此时距离最近的选择排在第一位；还可按好友间的亲密程度排序，此时与你联系最密切的选择排在第一位。我将这些设置比喻成“旋钮”或“滑动条”，可以将其价值调大或调小，就像在录制歌曲时用混音板更改不同耳机的输入平衡设置一样。

遗憾的是，数据服务商的旋钮并非总对用户可见。这些设置为什么如此频繁地采用黑箱化处理方式，让用户不可见呢？这并不是懒惰或贪婪等常见原因所致，也不是追求界面的简洁设计所致，其原因是经营方面的，例如创建旋钮需要花钱，可能会暴露所有权方面的秘密，或者某些旋钮可能会给公司带来法律风险。数据服务商并不从事定制程序设计工作，无论用户多么渴望，它们都不打算添加旋钮。但是，数据服务商应当赋予用户使用与调整现有旋钮设置的权利。此外，不设置旋钮也有认知方面的原因，让用户立刻明白旋钮所起的作用很难做到。即便如此，如维克托·迈尔·舍恩伯格（Viktor Mayer-Schonbe）与肯尼思·库克耶（Kenneth Cukers）在其著作《大数据时代》中指出的，预测与建议可以在人们未意识到它们工作原理的情况下发生作用。这些虽然属于法律方面的考虑，但也是限制用户主动性的借口。我认为只有通过使用旋钮，用户才能了解旋钮的功能和意义，从而选择最适合自己的参数设

置。

通过对旋钮参数设置开展实验，用户可以对数据服务商的工作原理建立心理模型。如果某个旋钮对最近数据和较早数据所分配的权重不同，用户就可以对此进行调节，以观察其如何影响推送的信息。与电源开关相比，旋钮通过更加动态化的方式帮助用户理解数据服务商的作用。例如，伊利诺伊大学开发的FeedVis程序，可以帮用户了解脸谱网如何获利。

旅行搜索网站嬉芒网通过分析价格、转机次数、飞行时间等一系列因素进行航班痛苦程度排名。虽然这是朝向正确方向迈出的一步，但用户如果不喜欢按照痛苦程度排列搜索结果，就可以单纯按照价格、飞行时间或转机次数搜索航班，还可以尝试根据具体的航空公司或旅行安排进行搜索（例如两次或两次以上的转机）。如果用户能对分配给各个痛苦因素的权重开展实验，效率将会更高。这可能也是嬉芒网感兴趣的方面。如果用户分配的权重与该网站的默认设置不一致，该网站就无法为用户提供较大的回报。例如，可将类似航线放在一起，以免用户在面对过多的相似选择时不知所措，可将搜索结果以时间轴的方式排列，直观地显示飞行时间。应当让用户对设置展开实验，并找到他们真正想要的航班。

以亚马逊收集的数据类型为例，该公司记录了客户的每一次购买行为和收货地址。除了根据客户的点击和购买数据提供个性化推荐外，亚马逊还会分析曾考虑购买此商品的其他客户与你的住所之间的距离。你想获得的商品推荐是你所在城市或州的人常会购买的商品。例如，如果你居住在经常发生旱灾的加利福尼亚州，你就希望获得关于更加节水的电器装置的信息。亚马逊还收集了其他类型的数据，它们可能也值得用户调整参数设置，例如在浏览网页时所使用的设备。在使用手机浏览网页时，如果能对手机下单的产品赋予比电脑下单的产品更大的权重，你就能更快地找到你需要的产品。这可能需要将设备识别指纹加入背景数据，例如，你现在正通过航班上的无线网络连接互联网就是背景数据。

数据服务商可能认为，公开自己的旋钮（包括相应的默认设置）会削弱它的竞争优势。实际上，某些组织之所以不愿赋予用户对数据开展实验的权利，是因为它们的业务收入中的很大一部分都源于信息不透明。几年前，22岁的实业家阿克塔尔·扎曼（Aktarer Zaman）创立了Skiplagged航班搜索网站，它可以发现两个城市之间低于航空公司官方

价格的机票。该网站利用的是航空公司有时推出的多航段打折机票信息，且航班中途必须经过非常热闹的枢纽机场。例如，某一天我想从旧金山飞往丹佛，我能搜索到的最便宜的机票是750美元。而同一家航空公司当天从旧金山飞往凤凰城并经停丹佛的航班，票价仅为500美元，更重要的是，这趟航班的第一航段正好是我要乘坐的航线。由于我没有乘坐这趟航线全长600英里的第二航段，我只支付了票价的50%，因为飞往凤凰城的人要比飞往丹佛的人少。扎曼称其为“隐藏城市”订票策略。美国联合航空公司起诉这家网站进行“不公平竞争”，因为它导致美联航收入管理系统中的某个副产品在客户面前变得透明。当数据服务商的利益可能与我们的利益相冲突时，通过了解这些旋钮（而不仅仅是数据），我们就能够发现并揭露类似的情况。

最重要的是，就像我们能获得工具对数据服务商的参数设置开展实验，并查看其对搜索结果显示顺序的影响一样，我们也能了解个人偏好的工作原理：当我们考虑不同的结果时的心理感受，以及这对我们的决策会产生什么样的影响。这是心理学家丹尼尔·卡尼曼与阿莫斯·特沃斯基的研究成果。他们针对不确定状况下的决策开展了一项划时代的研究，发现人们经常借助启发法或简单的心理规则，寻找令其满意的问题解决方法。卡尼曼与特沃斯基注意到三种常见的启发法：可得性启发法，它指想出某个办法或事物的容易程度；替代性启发法，它指对某个更具替代性的事物分配更高权重的意愿；锚定效应，它指相对于基准线判断事物的倾向。自从他们在大约半个世纪前发表这篇开山之作以来，对此问题的研究层出不穷，目前针对启发法这一主题已有几百种不同的版本。但是，只有通过实验，我们才能更好地了解这些启发法是如何影响人们的行为和决策。

让我们举个实际的例子。很难计算人们需要攒多少钱才能过好退休生活，因为在计算过程中存在许多不确定的变量。今后5年或10年的经济形势如何？能源价格的上涨情况如何？未来能发现哪些类型的新能源？今后的医疗水平如何？所有这些因素都会对未来的退休人士产生影响，但人们却对这些因素知之甚少。即便上帝能回答出所有这些问题，人们也无法对其产生影响。但是，人们能做到的是假设各种不同的情况，并针对自己能掌控的决定观察它会如何影响结果。通过调整该模型的参数，能让人们了解各种结果发生的可能性，这可能让他们更愿意选择与自己最初锚定的方案相距甚远的选项。我们应有权要求数据服务商创建假设并分享此类分析工具。

假设分析在人生的许多领域都十分有用。假设你是一名高三学生，此时你收到了哈佛大学和斯坦福大学两家名校的录取通知书。你如何做出选择？2014年，领英网推出了大学网站页面服务，借助网站自身巨大的简历数据库，分析大学毕业生的就业情况与他们的职业发展路径。经过挖掘的数据有利于假设分析和做出决策，而且重点既可以是初始情况（考虑选哪一所大学）也可以是结果（考虑选择什么职业）。如果你已经决定了未来打算从事的职业（例如，毕业后去谷歌、麦肯锡、孟山都或世界野生动物基金会工作），就可以查看哪一所大学的毕业生在这些公司的录用率更高。你可以查看某个行业主要是由哪些大学输送人才，包括某些十分吸引人的职业，比如非政府组织管理咨询、电视剧本写作或陶瓷工程。你可以对筛选功能开展实验，以便分析得出最适合你的专业，以提高你去某个公司应聘的成功率。

与许多决定一样，退休规划与选择大学都需要进行取舍。但是，人们经常在不得不放弃某个事物时，才知道自己对它的渴望程度有多大。通过考察取舍过程，我们将会逐渐明白自己希望或不太希望产生的结果是什么。通过对旋钮开展实验，人们能更好地提前了解自己的取舍过程。开展实验的权利为我们的决策开辟了新的信息渠道，使我们能更好地了解决策过程。

拥有自主导入和导出数据的权利

与修正数据的权利、对数据进行模糊处理的权利、对数据开展实验的权利一样，自主导入和导出数据的权利也旨在提高用户的主动性。在上一章中，我提到访问数据的权利不仅限于查看自己的数据，还包括查看对自己十分重要的数据。为了使透明性具有意义，你必须有能力解读你自己的数据。有了访问数据的权利，你可以要求数据服务商向你提供你的数据的副本，但在大多数情况下，除非你接受过另一个数据服务商服务，否则就无法发挥这个数据副本的作用。为了使主动性具有意义，你需要有自由地使用你自己的数据的权利——自主选择分享数据的范围和获得数据的对象。这就是自主导入和导出数据权利的基本目的。

在导入或导出数据时，你必须把数据从原先的位置移动到目的地。导入或导出数据后，数据依然存在于最初的创建位置。以人们熟悉的情况为例，某名学生希望将自己的大学成绩分享给一些研究生培养项目或

未来的雇主。即便他的成绩单已邮寄给这些组织，他的成绩记录仍然保存在他就读的大学里。这个例子虽然简单，但能吸引人们关注将导入、导出的数据用于其他地方所产生的影响。首先，收到成绩单的那些组织需要验证这份成绩单是否来自该生就读的大学，成绩单是否被篡改过。其次，该生可能只想把这份成绩单寄给自己选择的组织。他可能考虑过将大学成绩单发送给那些组织是否符合自己的利益。这样，他还可以通过信件或在面试时解释自己成绩差的原因，这就是对记录进行修正。将成绩单复印件寄给需要看成绩单的组织之后，该生便失去了对数据审查过程的影响。

长久以来，这一过程都是由人工完成的，学生将成绩单放入信封密封后，再将其寄给他选择的那些组织。在这个案例中，因为数据量相对较小，人工方式是有效的；每年的研究生培养项目大约会录取100万名新生，要求针对较长的时间跨度提供少量数据（对大学4年的学习情况进行简要概述）。

如果仅通过点击或移动鼠标，就可将10亿人创建的海量数据导入、导出，这需要更加复杂的技术解决方案。此外，当数据服务商的某项主要服务是分析并评价信誉时，它就需要格外小心，以防人们导入某些无法验证可靠性的数据。评级与审查系统（例如，电子港湾或亚马逊开发的此类系统）很容易被欺诈舞弊者导入虚假数据，以表明他们在其他网站上与客户有一流的交易记录。如果用户不再信任数据服务商提供的信誉数据，数据服务商乃至整个生态系统就会遭遇信任危机。同样，导出的数据也需要经过验证方可使用。

验证可通过加密密钥实现，许多人已经将它用于锁定或打开电子通信记录。你可以获得一对密钥：一把私人密钥不能和任何人分享，另一把公共密钥可以向任何人公开。假设你向某人发送一条信息，收信人需要确定这条信息确实由你发出。此时，你可以把自己的数据用私人密钥加密，再由收信人用你的公共密钥验证该信息确实由你发出，然后将此消息解锁并阅读。这种密钥组合的方法还可以解决另一个问题——你想给某人发送一条信息，并确保其他人无法阅读。你需要用收信人的公共密钥对数据加密，然后只有用收件人的私人密钥才可以解锁这条信息。将这两种方法结合，既可验证发信人的身份，也可限制阅读此条信息的对象。上述加密方法应当用于各种数据的导入、导出过程。

大学成绩单可以逐份打印并邮寄，数据的导入、导出也能逐份发送

到他们的个人邮箱中。这种打印并邮寄的方式非常适用于寄送大学成绩单，但如果社交平台上的10亿活跃用户申请获取其点击与移动鼠标的相关数据时，这种方式就无法胜任了。经过验证无误的数据还必须按照一定的格式发送，以便能直接导入接收数据的数据服务商的数据库。令人欣慰的是，目前已制定了数据分享协议，它就是应用程序编程接口（API）。它允许开发人员自主访问数据服务商的数据，无须向数据服务商提交一系列申请，也不必对所有结果逐条进行转换。应用程序编程接口使类似于嬉芒网等旅行搜索网站所提供的服务，通过它用户能在几秒钟内访问数十家航空公司的航班与票价信息。应用程序界面有助于开发人员未来对多个渠道的数据进行合并，推出新的产品与服务。

当应用程序编程接口向数据服务商发出“呼叫”时，数据在此刻相当于一张快照。例如，你在嬉芒网上进行搜索时，航空公司的航班和票价都是动态信息，但该网站不会不停地对搜索结果进行动态排序。当你还在对各种选择进行权衡时，某个航班座位可能已经售光了。如果搜索结果在你眼前不断地变化，你可能很难做出订票决定，因为你需要不停地重新考虑各种选择。

你的数据不应被锁在数据服务商的数据库内。如果你能将该数据与其他来源的数据进行合并、比较、对比，你就能从自己的数据中获得更大的价值。如果你创建出的某些重要的社交数据涉及可信度与声望时，这一点将尤为显著。优步与来福车等约车平台依靠对用户的评级和审核，使用户对其服务树立信心。审核与评级既适用于司机，也适用于乘客。司机的平均评级是评估其客户服务质量的关键性指标。2015年，如果优步司机的平均评级下降至4.6分以下（5分为满分），他就会面临账户被冻结的风险。另一个重要指标是司机的接单率。每一个客户的约车请求都会根据最佳的位置匹配原则按顺序发送给司机，每名司机都有约15秒的时间考虑是否接单。如果他在这一短暂的时间内没有选择接单，约车请求就会被转给系统中下一个匹配的司机。如果司机的接单率下降到80%~90%，他就会受到警告。多次受到警告之后，他就会被该应用屏蔽一段时间。如果他对三个约车请求不接单回应，应用就会在大约10分钟内停止向其发送约车信息，因为这种情况表明司机无法提供服务，此时如果为他分配一个订单，发出约车请求的乘客就不得不等待很长的时间。如果某位司机对所有派给他的订单都做出接单选择，之后再将其统统取消，企图钻规则的漏洞，他很可能会受到账户被冻结的惩罚。

该平台还根据数据分析来鼓励司机的工作积极性。在约车高峰时段，每家约车平台都希望能有更多的司机接单，以避免乘客等待的时间过长，因此每家平台都为抓紧一切时间提供服务的司机提供奖励。来福车规定，司机只要每周完成60个约车订单，且对分配给他的订单保持不低于90%的接单率，就不再对他之后所完成的订单收取佣金，以此让平台更具吸引力。与此相似，为了有资格参与优步公司在2015年推出的每小时最低收入计划，司机不仅需要将接单率保持在80%或90%以上（视具体城市而异），还需要在参加该计划期间达到一定的平台登录时间，通常是每小时的在线时间不少于50分钟，尤其是高峰时段（例如早晚高峰与周末的深夜），而且每小时至少要完成一个约车订单。尽管优步并没有明确禁止司机加入另一个约车平台，但司机如果加入别的平台就会失去优步提供的最低收入保证。

这些补贴计划牢牢地吸引住司机。如果司机在某个约车平台上积累起优秀的信誉，他就必须做出选择：继续使用该平台，还是加入其他平台？但后者需要从头开始积累信誉。如果他的信誉数据被锁定在某个平台上，他接到约车订单的能力也会逐渐被锁定在该平台上。

自主导入与导出数据的权利挑战了这一现状，它将主动权从商业组织转移到个人手中。从事风险投资的合广公司的阿尔伯特·温格（Albert Wenger）认为，按需经济的劳动者（包括优步、来福车及其他约车公司的司机）应当拥有使用“应用程序界面密钥的权利”，从而能够通过应用程序编程接口访问他指定的数据。当用户与数据服务商就数据产品与服务讨价还价时，这一权利能使信息的使用更加公平。用户可以将自己的信息转移到新的“市场”，例如一家新的约车网站。如果司机的评级非常高，约车公司就会愿意向他们支付更高的价格来挽留他们。如果能将信誉数据、交易数据与其他数据从一家公司复制到另一家公司，就能帮按需经济的劳动者提高谈判筹码。自主导入与导出数据的权利可以确保用户的信誉始终跟随着用户，就像现实世界中的信誉一样。

此外，这一权利也迫使公司更关注创造好的产品与服务，而非只储存数据。针对面向客户的网络公司，如果我们现在回顾它们近20年的发展历程，就会发现这些公司的重点显然始终放在收集更多、更好的数据上，这些公司通常比努力改进算法的公司更成功。实际上，如果公司想要提高个性化服务水平，就有积极性去接收用户导出的数据，因为来自其他渠道的数据常能提升个性化服务的质量。

用户有权决定是否将自己的数据提供给允许用户导入和导出数据的数据服务商。用户应要求老牌的大型数据服务商（谷歌、脸谱网、亚马逊）允许用户导入、导出数据，因为它们相对于刚开始收集数据的新公司具有内在优势。从用户的角度看，导入、导出数据的权利确保用户的数据不会受到某个数据服务商的绑架，即便某些数据服务商能提供此功能，用户也能找到提供此功能的其他数据服务商。

1 000年以来，人们一直在努力争取人身自由迁徙的权利。我们现在还必须努力争取数据自由迁徙的权利，在这场数据革命中，流动性是实现人的主动性的关键。

人类擅长的事和机器擅长的事

通过这4项与主动性有关的权利，用户就能够掌控影响数据服务商产出的用户个人数据与参数设置，而非要求“有关部门”精确地规定数据公司对用户数据的使用时间、地点与方式。但是，明确人们相对于计算机的优势是什么，很重要。我认为，我们应当让人们去做人类擅长的工作，让计算机去做计算机擅长的工作，而非将两者混为一谈。

为了方便读者理解，我将以一个早期技术发展的例子说明为何人们要将自己的某些控制权让渡给机器。20世纪60年代，几家大型汽车制造商与工程公司探索了在汽车上使用防锁死刹车系统（ABS）的可能性。飞机上已经采用了类似的系统，它让飞行更加安全。如果飞行员在制动时犯错，将会危及数百条人命，因此航空公司极为迫切地想通过该系统降低事故的发生概率。但是，销售代表、客户，甚至某些业内专家对此仍然持怀疑态度。他们声称：“客户绝不会允许由一堆晶体管来做出刹车的决定，因为这事关他们的生死。”1978年，博世公司生产出第一款标准的防锁死刹车系统，并将其用于梅赛德斯–奔驰与宝马的顶级车型。通过反复的安全测试，证明计算机在车辆制动时能够更加可靠、精确地控制刹车装置，帮助驾驶员更好地控制车辆的方向。如果机器和人类可以协作，就能让驾驶更加安全。根据数十年来收集的证据，多个国家的政府认定不安装防锁死刹车系统的汽车缺乏安全性。如今，美国与欧盟生产的每一辆新车都安装了防锁死刹车系统。

自从防锁死刹车系统研发以来，我们将越来越多的驾驶任务委托给

车载电脑完成，对人类来说驾驶变得更轻松。但我们还未将所有的决策权都交给机器，机器只负责完成比人类更擅长的任务。以“定速巡航控制”为例，早期的定速巡航系统能够让汽车保持匀速行驶，无须司机控制油门。但是，计算机尚未学习接收与环境有关的各种数据，也不能为根据具体环境对速度进行调整。因此，这些依然由司机根据自己对当地法规与道路状况的评估独立完成。

更加现代化的“适应性”巡航控制系统，已经能让计算机在分析环境方面发挥一定作用了。某些系统能根据传感器数据对司机发出警告或改变行驶速度。例如，如果探测到其他汽车或障碍物与自身车辆距离过近时，系统就会让车辆减速。这确保汽车能通过减速保持“安全”距离，而不是由司机先意识到危险，再在即将发生危险前关闭定速巡航控制。通过交通信号灯识别与警报系统，宝马、梅赛德斯–奔驰及其他汽车生产商将来自红外距离传感器的数据与挡风玻璃上的摄像头的数据相结合，再通过图像识别软件进行处理。在某些情况下，还要与带有地理位置标记的法定限速数据库进行对比。自动制动系统（AEB）是驾驶过程中人机分工的下一个发展阶段。在初期研究中，自动制动系统能减少近1/2的追尾事故；汽车前向碰撞预警系统则能减少约1/4的追尾事故（它向司机发出警报，但没有制动功能）。还有更先进的技术能够让计算机之间直接进行通信，无论车辆是以目标速度行驶、变换车道还是寻找停车场，都能在司机不知不觉中与其他车辆交换位置、速度、行驶方向与目的地方面的信息。驾驶领域的每一次创新，都是我们实现自动驾驶过程中的里程碑。

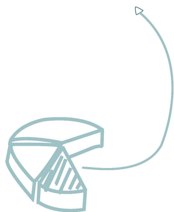
如今，计算机在驾驶安全方面担负着重要责任。当你坐在方向盘前驾驶汽车，却发现定速巡航控制系统正在让你的车辆加速（因为传感器探测到以更快的车速行驶是安全的）时，你会愿意吗？计算机能发现司机的偏好并调整相应的参数，但我们需要依据自己的决定可能产生的结果设置权重。社交数据将越来越多地与商业、金融、工作、教育、医疗、政府管理结合在一起，用户应当要求数据服务商提供工具让其变得更加透明，并提升用户的主动性，保持对重要问题的掌控力。我们必须了解数据如何影响我们的决定，并有权独立做出这些决定。



第7章
把未来创造出来

行使人类对数据的权利

如何在生活中体验数据为我们带来的好处？



预测未来的最好方法，就是把未来创造出来。

——阿伦·凯（Alan Kay）

对权利的讨论很轻松，但如果这些权利无法对你的日常生活产生影响，它们就毫无意义。我们面对的问题是怎样使数据的使用成为可能并被广泛接受，这是一个重大问题。零售商在向你提供个性化服务时，能够使用哪些类型的数据呢？借贷方在决定是否批准你的贷款申请之前，有权查看你的脸谱网好友列表吗？或者说这是21世纪的“贷款歧视行为”（根据居住地点对贷款申请人区别对待）吗？当你把自己的健康数据提供给雇主的定点医疗机构时，你怎么知道它不会对这些数据进行分析，并将其用于与你本人或你的工作有关的其他决策呢？通过进行收集学生的数据，我们就能对课堂设计进行优化，以便真正确保不会“落下一个孩子”吗？分享数据有助于我们更好地做出决定，而且使决策过程更加智能化。但是，我们应当尽可能地了解分享数据可能给人们带来的好处与坏处，还要了解如何运用与透明性和主动性相关的权利。

按照你自己的需求购买产品与服务

人们在购物时，通常会对需要购买的产品与服务的价格、规格、评级、用户评价等货比三家。社交数据极大地减少了传统的信息不对称程度，而且客户购买模式的透明化也会改善人们的购物决策过程。以亚马逊为例，我们希望了解哪些类型的数据对人们的购物决策过程帮助最大，答案是浏览数据（“浏览了这件商品的客户，也浏览了……”）与购买数据（“购买了这件商品的客户，也购买了……”），或两者兼有。我们发现，如果客户能获得其他客户的点击与购买情况信息，就能提高客户的满意度。

传感器数据也有助于公司提高产品的透明性。以总部设在新西兰的美利奴公司为例，该公司生产的每一件服装都附带一个独一无二的条形码，由数字与字母混合组成。客户将所购买服装的条形码输入该公司的网站后，就能追溯到这件服装所用羊毛纤维原料的产地，还能浏览具体的牧场信息。我在该网站上查询了我买的羊毛衫，得知它的原料产自布朗奇奇克牧场，该牧场占地超过1.6万英亩[注]，雷·安德森（Ray Anderson）在那里养了9 000只绵羊。我还了解到，自从雷的爷爷在第一次世界大战后退伍还乡，他的家族就一直在经营这座牧场。

对服装原料的来源进行介绍是一种温馨的营销方式。该条形码还有另外一个用途，它有助于美利奴公司追踪在全球哪个地方会出现假冒与仿冒该公司的产品。造假分子得知客户会验证条形码，就会立刻在伪劣产品上使用类似的条形码，以便在外观上能做到以假乱真。但如果客户将此条形码输入网站，并发现此条形码系多次使用或无效时，美利奴公司的数据科学家就能展开追踪调查工作，圈定假冒伪劣产品可能的位置，以便公司告知供应商与零售商这一问题。过去，假货只需尽量模仿品牌产品的外观就可以确保它们不被发现；但现在即便外观相同，也依然会被发现。

我预计未来许多产品（以及产品的零部件）都会附带独一无二的识别码，它可以是条形码、二维码或无线射频识别（RFID）标签。许多公司正在探索产品追溯的方法，追溯过程可以从生产的初始阶段开始；在食品包装上也可以贴上这种标签。二维码可以通过手机摄像头扫描识别，商家已将其用于追溯鲜鱼是从哪个码头送到了哪个批发市场，以便厨师能照顾当地渔民的生意。它们还被用于验证药店的某种瓶装药是真

是假，这是打击假药的一项有力举措。此外，识别码数据也可以与生产过程中每个阶段的传感器数据相结合，让客户知道食品或药品始终存储在安全的环境中。只有在公司网站上有关于产品原材料的来源地与组装方式的显性数据，客户才会从该公司购买产品。有一家公司名叫应用DNA科学公司，它出售液态DNA，液态DNA是从植物中提取并重新组合的独一无二的基因字符串，它可涂在产品表面并在中央数据库中注册。如果执法人员找到了丢失的产品，就可以通过化验了解其独一无二的DNA，并将其归还给合法的所有者。由于液态DNA的用量极小，因此假药鉴定也在试用这项技术。

针对高透明性的公司推出的可溯源产品或不可溯源的产品，客户会在两者之间做出选择。实现透明性需要付出代价。但缺乏透明性也需要付出代价。无论某件产品是否需要遵守政府制定的法律，你都应当能访问该产品的相关信息，例如原材料的产地、生产线的状况、上架之前经历了哪些环节。与此相似，在你决定对自己购买和使用某件产品的数据进行分享时（包括对自己购买的产品进行注册以获得质量保证），你都应当能看到公司是如何使用这些数据的。如果公司告诉你该产品存在缺陷需要召回，或者它将帮你找回失窃的物品，你就会认为这是公平的交易。如果该公司承诺这些数据仅用于向你推荐你真正感兴趣的产品，而非由营销团队推送的产品时，你可能认为这也不错。但如果公司仅把这些数据用于向你发送垃圾广告，你可能会讨厌这种交易，更不要说公司将你的数据出售给他人，导致你成为后者的营销目标了。

将可见性加入产品的寿命周期（销售前、销售中、使用中、使用后），可以改变消费模式。如果公司能通过独一无二的识别码跟踪产品的整个使用过程，它就能为客户提供回报，例如推荐更适合这些客户使用的产品，或对参加此产品回收利用计划的客户提供奖励。麻省理工学院的垃圾跟踪项目通过功能十分有限的手机来完成，研究人员将其附在垃圾上以查看最终进入回收中心与垃圾填埋场的垃圾分别有多少。这些手机的程序设定为每天“自动唤醒”几分钟，探测垃圾当前的位置，再编辑文本信息发送到研究人员的中央服务器中。该项目的一个目标是了解罚金与补贴对某个社区垃圾回收率的影响，另一个目标是通过提高垃圾的可见性提升垃圾的回收率。

我在之前的章节中指出，你的移动电话公司必然知道你去过的地方与你拨打电话的对象。电信公司有时会打着节省话费的幌子，根据你的

语音、短信、数据的使用情况，推荐你使用另一个套餐，但这是通过使用你的数据后向你提供的服务。此外，电信公司还可以利用这些数据推出十分有用的服务。

以亚历克斯·阿尔加德的Hiya公司与骚扰电话过滤器为例，它会提醒你，你接到的电话很可能是推销电话或骚扰电话。与云技术电话服务商Skydeck公司相似，它也提供“好友关系管理”服务，通过对你通话模式的分析发现你可能会与某个好友失去联系，并向你发出警告。电信公司甚至能根据电话的使用情况提供健康警告。如果某人处于临床抑郁症的急性发作期，他的手机使用情况和行为模式就会发生变化，这可以通过他的位置发现，例如他很少外出，上班或参加其他常规活动的时间不规律，而且会花更多的时间在手机上，但并非用它来打电话。他可以要求电信公司在这些情况发生后通知他，帮助他了解自己的心理状态，或者要求电信公司给他的某个值得信赖的朋友或医生发送相关信息，让他们来看他。数据回报评分有利于在这些领域及其他领域推动创新。与主要从事普通电话服务的电信公司相比，能为客户提供额外的好处、提供新型服务的电信公司将获得更高的数据回报评分。

这些例子表明，访问和检查数据的权利有助于客户做出更好的决定。但是，客户获得修正与导入、导出数据的权利之后，可以获得更大的权利。因为这些权利还让他有机会发现更好的匹配，包括更能满足其个性化的购买需求。为了考察这一过程，让我们以航空旅行为例。

几十年来，客户对航空公司的票价制定方式几乎一无所知。价格、日期、航班号与其他细节都是固定的；纸质机票与登机联均需打印，登机联要提供给登机口检票人员才能上飞机。一旦出票之后，改签机票的手续比较耗时，而且收费较高。

现在，航空公司的时刻表与收入管理均已实现电子化，几乎所有的机票也都采用电子机票形式。自天巡网、客涯网、嬉芒网等在线旅行服务商出现之后，客户就能够对各家航空公司与各条航线的票价进行对比，以便查看票价的变动方式并预计票价在短期内的变化情况。这些网站能对成千上万个航班进行分类，并对旅行的其他方面（例如，飞行距离或时间）的权重分配方案开展实验，例如嬉芒网的痛苦度或类似的服务。航空公司甚至找到了利用旅客对票价变动的焦虑情绪的方法——旅客可以支付一笔定金以便在几天内锁定票价。选择各种排序方式归根结底都是在寻找最适合自己的票价。由于航空公司的机票已不再是纸质机

票，而是电子机票，我们也不再有一成不变的购票需求了。

修正数据的权利可以为机票的新型购买与销售模式奠定基础。假设你能根据自己的旅行计划的弹性情况更改机票。例设，你提前购买了一张价格为350美元的机票，这是当天从波士顿飞往旧金山的第一趟航班，它于早晨6点起飞。但你当天的出行计划实际上弹性很大，你可以备注机票，以便让航空公司知道你愿意乘坐早晨6点之后起飞的航班，但条件是能获得200美元的返现奖励。几周之后，另一名旅客在线预订机票。他想乘坐早晨6点钟的航班，以便能赶在午餐前拜访旧金山的一位朋友，之后再去参加工作午餐会，但当天早上6点与7点的航班均已售罄。他只能购买下一趟航班，并备注他愿意额外支付300美元乘坐早上6点的航班。航空公司的售票系统发现这一匹配之后，就会将你的座位分配给这位旅客，并对他加收300美元，然后帮你改签至之后的航班并向你支付200美元，其余100美元则成为航空公司的额外收入。对已购机票的备注，必须具有与机票同等的法律约束力。在某些情况下，你还需要设定备注的使用时间与方式。

如果常旅客计划为客户提供数据权利，就会影响客户与航空公司之间的关系，因为大量的客户忠诚计划都旨在将客户牢牢地留住。

假设你住在达拉斯，达拉斯机场是美国航空公司的枢纽机场之一，而你经常乘坐该公司的航班；再假设你在美国航空公司的常旅客计划中已升级为金卡客户。你到了美国联合航空公司的枢纽机场——休斯敦机场之后，你在美国航空公司的金卡客户身份（包括使用机场贵宾休息厅、提前值机与免费升舱等特权）在你乘坐美联航的航班时却毫无用处。如果你能将自己的金卡身份、有效期限，甚至你乘坐航班的所有历史数据导出到美联航的系统中，美联航就可以选择为你提供与美国航空公司相同的待遇，争取让你成为它的客户。

毫无疑问，将本书提出的与透明性和主动性相关的权利提供给客户后，不仅能使他们更乐于将数据分享给公司，而且它会让客户关系管理发生逆转。公司成立的目的是通过更加灵活地向客户提供新的产品与服务，从创新思维中获利。但最重要的一点是，通过让数据服务于人，使客户获得权利。

金融的未来

行使数据权利不仅会改变人们的消费方式，而且会改变人们管理消费的方式。过去，小镇上的信贷主管常去了解每个人的贷款状况，这是乔治–贝里公司的经营风格。如今，你很可能需要向一家大到不能倒的跨国公司申请评估自己的信贷价值。在你申请贷款时，你可能不希望信贷主管检查你的详细交易历史，或脸谱网时间轴，以免发现你的不良行为。在现有的海量数据中，许多数据可以在你向银行申请贷款时给你提供帮助。你可以自行决定想让银行看到（或看不到）的金融数据，但如果你选择不向银行分享其认为必要的某些数据，你就必须接受无法获得贷款的结果。无论如何，分享多种个人数据对于信贷历史有限的人极为重要，比如刚刚走出校门的年轻人。

以刚从大学毕业的米格尔为例，他的背很疼，急需买新的床垫。米格尔手头很紧，拿不出几百美元的现金，他的信用卡额度也已用完，他知道超出透支额度的欠款利息高达39.9%，就算买到了床垫也无法让他睡个好觉。于是，他在凯斯柏网（一家网上床垫零售商）上结账时，选择“通过Affirm支付”，这是金融科技初创公司Affirm提供的短期贷款选项。他从屏幕上看到选择3期、6期或12期分期还款计划，每个月分别需要偿还的金额。米格尔选择了其中一种分期还款计划并通过Affirm进行了支付，凯斯柏网站立刻获得了全款，并将床垫送到米格尔的家中。在此交易中，米格尔既不需要一下子支付床垫的全款，也不需要担心每月产生的利息。

Affirm公司的首席执行官麦克斯·拉夫琴（Max Levchin）是贝宝公司（PayPal）的合伙创始人，曾担任该公司的首席技术官。他希望通过Affirm彻底改变消费者的信贷行为，就像贝宝彻底改变了在线支付一样。他认为，社交数据可以让更多的人获得贷款。该公司利用5类以上的信息对米格尔进行金融信用评分，从而为缺乏信贷历史的人群更好地评估信贷风险。这些信息包括网络浏览行为、脸谱网与推特上的活动、手机呼叫与短信发送频率，以及手机的操作系统。该公司还考察贷款申请人在GitHub等在线社区中的活跃程度。该社区是软件开发人员分享自己的代码并与他人进行分工合作的地方，该网站的发帖人通常都有认证身份，他们的成果也会获得声望反馈。对于某些贷款申请人，Affirm公司会要求临时检查其支票账户，以分析他们的购买与收入模式。

有些金融科技初创公司为“无法充分利用银行服务的人群”提供服务。金融初创公司Upstart致力于向二三十岁的人群提供贷款用于偿还信

用卡。该公司并不完全依据对你的当前收入与开支情况的评估发放贷款，还要审核你所上的大学、所选专业、所学课程、成绩、高考分数，预测你在未来几年内的加薪趋势，从而计算出你偿还贷款的能力。与此相似，金融科技初创公司ZestFinance声称自己收集了许多个数据点，以确定何时对低收入人群和没有开立银行账户的人群批准信贷展期。数据科学家发现，如果贷款人在填写贷款申请表时全部使用大写字母，他们偿还贷款的能力就很有可能低于同时使用大写与小写字母的人。

在这些案例中，金融科技公司正在利用社交数据确定是否应向信用记录糟糕或信贷历史有限的人群放贷。截至2015年，每5个成年中国人中就有1个人获得了金融信用评分，社交数据在消费信贷领域中已经开始发挥着关键作用。2016年，中国完成了大约2亿张信用卡的申请。审核通过率大约为30%，申请被拒的原因不仅是缺乏信贷历史或信贷记录糟糕，还因为中国政府对每家银行可发放的贷款金额有所限制。

中国政府向芝麻信用发放了信用评级牌照，这是阿里巴巴公司开发的示范项目。每年有超过6.5亿人使用阿里巴巴公司的淘宝网，这为芝麻信用访问海量的交易与沟通数据提供了机会。据报道称，阿里巴巴公司通过它的支付系统——支付宝，在2015年11月11日（简称“双十一”，这是阿里巴巴为消费者创造的“购物节”）创下了140亿美元的单日销售额纪录。其中大约有70%的销售是通过智能手机完成的。阿里巴巴公司可以查看支付宝记录下的地理位置数据，以识别买家在哪里过这个“购物节”。人们在聚餐时，支付宝应用可以提供对应的支付选择，即“AA制”。这使阿里巴巴公司获得了真实世界的信息，不仅知道人们点了哪些食物，而且知道他们与谁一起用餐，以此计算芝麻信用得分。

交易数据与社交图片数据对金融机构做出是否批准信贷申请的决定越来越重要，因此人们也应有权访问这些数据，并了解它们会如何影响你的信用分，就像了解是否按时还款在金融信用评分中所占的百分比一样。通过分析你的位置数据，可发现你利用时间的方式。如果你大部分时间都待在办公室，你的贷款申请就很容易得到批准；如果你有很多个夜晚都在酒吧消遣，就会降低你的贷款申请被批准的可能性。数据服务商会分析你的社交图谱，查看你的某些朋友是否属于信贷高风险人士，这类似于好事达保险公司的做法。比如在案件调查中，该公司发现你的某些朋友曾经有骗保行为。如果你与这些人来往甚密，就可能导致借贷方拒绝你的贷款申请，当然你有权知道他们是谁。就像脸谱网会告诉

你，你在哪些照片中出现了，银行也应当告知你，它在信贷审批过程中使用了关于你的哪些数据。

在你审核完这些数据之后，就要立刻决定应改变自己的行为，还是通过修正或模糊处理改变数据。你可以通过备释当时的背景来修正数据，就像对大学成绩单上得分较差的科目做出解释一样。过去几十年来，人们在当地发放购房贷款的机构接受信贷主管的面试时，就是用这种方法解释自己为何未能正常还款的。在看到你的社交图谱中的每个人对你的贷款申请所产生的影响之后，你可以决定与拉低你的信用分的人解除好友关系，就像小镇上的人会与声名狼藉的人断绝往来一样。如果数据服务商将你在社交网络中的好友的好友的信用情况也考虑在内，结果可能非常复杂。这就是本书提到的脸谱网的案例，即该公司在多个领域使用社交图谱的专利，其中也包括金融领域。你可能会在提出信贷申请时同意金融机构审查你的脸谱网社交图谱，但前提是你有权对图谱中好友的信息进行模糊处理。

在某些情况下，你的社交网络可能有利于你获得贷款，因为你的好友都是按时还款的高信用等级人士或他们愿意为你贷款中的一部分金额履行有效的担保责任。你可以通过对自己的社交图谱开展实验的权利，选择将哪些数据导出给借贷方。

导入、导出数据的权利不仅对于你的贷款申请极为重要。当你想进行投资时，导入、导出数据的权利可以让投资条款对你而言更公平。投资领域中，这方面的先驱者是SigFig财富管理公司，它由迈克·沙（Mike Sha）创建。我在亚马逊公司工作时，他负责管理亚马逊的支付产品。他离开亚马逊之后，打算通过对个人投资者在整个经纪行业产生的数据进行汇总与分析，帮助改善他们的投资决定。

不久之前，金融经纪人通过提供咨询建议受到人们的信任和尊敬。金融经纪人会与他们的客户面对面坐在一起，努力了解他们的长期金融目标并向他们推荐投资组合。他们很少会透露对这些服务的收费情况，因此客户的谈判筹码很少。在迈克看来，金融经纪人显然不可能透露法律规定之外的内容。迈克告诉美国知名科技网站Business Znsider：“你无法电话询问他们的收费情况，我们和金融经纪人交谈过，他们也不清楚其他公司的收费结构。”单个经纪人的绩效数据尤其难以获得，他们通常是巧舌如簧的推销专家，但却无法提供透明性。个人投资者怎样才能获取所需的信息，以便做出出色的投资决策呢？迈克的想法是采用“交

换”的原则解决问题。

迈克要求客户向SigFig财富管理公司提供对其账户的只读访问权利，以换取全面的投资分析方案和财务状况评估。SigFig软件会检测出缺乏效率的地方，并提出改进投资组合的方法。即便经纪商不支持客户数据的导入、导出，该公司也能通过模拟用户登录账户对其数据进行屏幕抓取处理，当然这需要获得用户的允许。这相当于通过与经纪商合作导出客户数据，但这样做的影响力同样巨大。100多家金融机构的80多万名客户已授权SigFig财富管理公司访问其账户，涉及总金额高达3 500多亿美元。巨大的数据量使该公司能够向用户提供前所未有的透明性，这是任何一家经纪商都无法独立提供的，也是它们不愿提供的。SigFig财富管理公司计算出投资者实际支付的费用，并将此信息分享给投资者。它还分析出每个投资者的投资组合绩效，并将其与交易型开放式指数基金（ETF）及其他低成本投资产品进行对比。这些信息使客户能够与其经纪商进行谈判，以获得更好的交易或服务，否则就会把资金转到它的竞争对手那里。SigFig财富管理公司通过为个人投资者提供导出数据的手段，迫使经纪商实现透明性。

在中国，个人客户也在利用技术工具带来的优势，提高与金融决策相关的透明性，包括信用卡申请过程。由于中国政府对每家银行分配的信贷额度是不公开的，因此该过程十分复杂。涂志云曾在金融信用评分公司和安客诚公司任职，回中国后他创立了我爱卡网站，指导信用卡申请人完成缺乏透明性的审批流程。与SigFig财富管理公司一样，人们必须允许我爱卡网站获得他们的个人数据，在此情况下，这些数据是申请信用卡通常所需的数据。通过分析这些数据并将其与之前获得批准或被拒绝的申请进行对比，该网站能够指导人们去最有可能成功获得审批的某些银行办理信用卡。如果银行不对其客户提供透明性与主动性，人们就会寻找工具撬开它。

公平的职场

在担任德累斯登佳华投资银行的首席信息官期间，J·P·兰加斯瓦米（J. P. Rangaswami）认识到一个严重问题：处理雇员的抱怨占用了他太多时间。他的电子邮箱中满是雇员对其他部门、上级主管、团队成员的投诉信。有些投诉是合理的，但也有很多纯属职场当中的钩心斗角。

兰加斯瓦米没有时间将所有信息区分为哪些需要处理，哪些不值得处理。作为首席信息官，他针对公司的局域网建立了某种制度，以便员工之间能在可信度与贡献方面相互进行评级，但这必然会对士气产生负面影响。在收到众多抱怨邮件后，他想出了一个更加简单的方法，他给予自己的直接下属访问自己收件箱与发件箱的权限。

他注意到，员工的投诉邮件数量立刻减少。当然，并非所有人都欢迎这一举措，某些员工在几个月后辞职。当得知其他人也会看到自己所发的电子邮件后，员工的行为方式发生了改变。之后，兰加斯瓦米很好奇这些人会如何查看他的邮箱，他想了解他们的想法，或者用他的话来说："要进入他们的内心世界。"他发现人们对他的发件箱比收件箱更感兴趣，这意味着他说的内容比别人告诉他的内容对他们而言更重要。

这些都是兰加斯瓦米在2001年采取的措施，三年后才出现了谷歌邮箱和脸谱网，10多年后才出现了Slack公司。Slack公司是一个职场沟通平台，它对公司中的每个人公开每一份备忘录与信息，雇员现在已经获得了对海量信息进行处理所需的强大数据修正工具。用户利用精心设计的评价系统回复主题和帖子，所有的主题和帖子在得到回复后就会向上浮动。评价系统中包括表情符号，这比单纯给帖子好评或差评，甚至更简单的点赞按钮更有表现力。未来，Slack等沟通平台可通过对发帖进行实时的语义分析，以及利用摄像头对用户脸部表情进行情绪识别，向用户推荐一些表情符号。

如今，公司可以以较低的成本衡量雇员的产出（每一次敲击键盘、每一帧视频），并分析他们的工作合格程度与绩效。假设某家公司让其雇员参加类似于社交计量标牌的计划，主管能够监控雇员间的互动情况，以及他们在不同环境下的产出情况。例如，让传感器对雇员警惕性、情绪和晚上的睡眠情况进行监控，这可能成为今后普遍的做法。通过访问这些数据，有助于雇员确定应当在何时、何地完成某些类型的工作。该系统可以向雇员推荐某些类型的工作，因为这更适合某人当前的状态。主管也可以根据这些数据，选择不让某个雇员参加某个大型发布会。如需了解参加此类数据收集与分析计划对自己是否有利，雇员需要拥有查看数据回报评分的权限，而且应该从雇员的角度而非从主管的角度打分。

就像公司正在探索新的数据源一样，雇员也应有权查看公司绩效评估和补贴的计算公式，包括有关数据及其各自权重的完整列表。这种透

明性有助于雇员更好把将自己的时间和精力用于公司目标的达成上。如果公司将多个来源的数据汇总，例如电子邮件与电话沟通模式，社交计量标牌的读数，同事的评价、评级与调查等，雇员就很难对自己的工作成果造假。通过分析这些数据输入，主管与雇员都能看到想法是怎样在组织内部传播的，发现非官方专业技能的聚集点与沟通瓶颈。拥有访问和导入、导出数据的权利，更易于雇员发现主管的行为何时与系统的提示有差异，表明主管可能存在偏见与歧视行为。

将就业数据分享给外部的数据服务商之后，我们还能更好地了解影响我们职业生涯的经济大趋势。领英网可以根据4亿多名用户所分享的数据，判断某些公司与行业是否健康并归纳出其特点。举一个惊人的例子，领英网的数据科学家注意到该网站在2008年9月14日（当天是周日）十分活跃。由于这在周末属于异常情况，他们担心网站受到黑客攻击。他们召集了安全团队，经过调查后，他们找到了数据流量的源头。所有这些行为都来自雷曼兄弟公司的雇员，他们疯狂地联系他人、更新简历、下载联系人信息。领英网怀疑这表明雷曼兄弟公司避免破产的努力宣告失败，但当时新闻尚未公开予以确认。

雇员下载各种联系人的信息是不好的征兆，雇员大批离职也是一个不祥之兆。如果某个公司的人才不断流失到同一行业的其他公司，它的前景可能就会比竞争对手黯淡。上述信息目前仅被提供给企业客户。领英网针对个人用户开设了大学排行页面，表明对于某个大学的毕业生来说最受欢迎的雇主是哪家公司。该网站针对从某家公司跳槽的员工，可以帮他们找出最受他们欢迎的雇主是哪家公司。这相当于亚马逊的商品推荐服务，它还能计算出之前在某家公司工作的员工中，目前在另一家公司就职的比例是多少。

社交数据不仅可以为某家公司的雇员结构进行优化，还能对他们的工作时间进行优化。例如在零售行业（及其他行业）中，为雇员排轮班表始终具有挑战性。许多变量都会影响在某一时间或某一天前往某家商店购物的人数，例如天气（寒流或倾盆大雨）和营销活动（热门的促销活动或广告活动）。田中·格雷格（Greg　Tanaka）是零售店分析服务公司贝深科技的创始人，他与他的团队建立的模型能对零售店的客流量进行预测，并确定接待客户需要多少名员工。对员工人数进行优化是关键。格雷格解释称：“每三位离开商店的顾客中，就有一位是因为他们找不到销售人员为自己提供服务。”但是，始终超额配置员工在经济上不具

可行性，因为零售业的利润空间十分有限。人与人之间的沟通不容忽视，团队合作的力量可能比所有当班个人的绩效总和更大。

虽然优秀的主管能敏锐地预知繁忙的销售时段，但这种基于个人观察所做的预测无法与贝深科技的模型相媲美，因为后者是根据该公司安装在零售店内的摄像头与话筒所收集数据建立的。视频与音频不仅能用于测算客户人数，还能衡量客户对产品的购买情况，既包括整体噪声水平，也包括哪些购物区域或产品最能吸引客户。这些数据有助于零售店了解客户购物的频率，以及他们通常倾向于在哪些区域购物。此外，贝深科技公司还能发现哪些工作人员共同工作时销售业绩最高，通过将他们组成团队，零售店能够在人员成本不变的前提下增加约10%的收入。

贝深科技公司的惊人之处在于它挑战了困难的任务——制定每天的轮班表。格雷格建议工作人员分享自己的日程安排，并在预测到某个班次需要增派人手时，简化临时通知工作人员的程序。工作人员可以对自己日程安排中事项的具体细节进行模糊处理，只需分享各个时间段是否空闲这一信息。但只有少数人参加了这项计划。原因何在？他们不愿意将此数据分享给他们的主管吗？格雷格与工作人员交谈后得知，原因很简单：许多工作人员都没有在线日程表，即使有，也很少对它进行更新。要想让他们创建有关其空闲时间的数据，必须向他们提供奖励。

修正数据的权利对此十分有效。某个工作人员上班的意愿通常并非只有愿意或不愿意这两种情况。工作人员可能愿意在最繁忙的时段工作，尤其在他们可以拿到奖金时。他可能会注明自己在某些时段空闲并可安排加班，如果某个班次需要加派人手，便可将他自动分配到该班次。如果他有空闲时间可以加班几个小时，他就可以对这一时段的数据进行修正，备注只在有奖金或提高工资标准的情况下才会同意加班。如果公司保持现有的工资水平不变，他还可以采用另一种修正方式，即获得一些积分。积分可被分配至下一周的某些时段，以表明他对在这些时段上班的偏好高于其他时段。主管可以给表现优秀的员工额外的积分奖励。在这两种情况下，工作人员都可获得更多的谈判筹码，主管则可实现更高的工作效率。通过提高工作人员的主动性，改善了整个职场的生态系统。

对工作数据开展实验的权利，有助于人们在职业生涯中取得进步。领英网的大学排行页面有助于学生对不同的教育情况展开实验，以查看每一种情况分别产生各种职业成就的概率。如果对这项服务予以拓展，

就能帮助你了解未来自己赚钱的能力、晋升时间、技能水平（通过分析简历或技能认证予以确认）、所在部门和公司。简历中的不同措辞会对搜索结果排序产生影响，如果能对此开展实验，就能帮助用户在就业市场中与雇主更好地进行匹配。

发现人们的才能需要时间与资金，这就是各家公司越来越多地依靠社交数据为他们实现这一目标的原因。数据战略公司MoData的首席执行官兼创始人甘姆·迪亚斯（Gam Dias）在招募数据科学家时，发现能够在在线问答网站Quora上不断给出高质量答案的发帖者很可能真正了解他参与讨论的问题。甘姆说："人们会登录网站，阅读帖子，发表帖子，对别人的发帖进行评论。在线问答网站Quora从事的是一种知识经济，我无法在发帖质量、人气、影响力方面找到能与之匹敌的网站。"他发现该网站的几位发帖者在机器学习领域中具有很高的声望，而且他特别喜欢某个长期发帖者的帖子。虽然甘姆联系这位发帖者并得知他对换工作不感兴趣，但这位发帖者还是同意飞往硅谷参加面试，最终加入了甘姆的公司。

10年前，大多数人都会认为在网上公开分享自己辛苦获得的专业技能，就是一个笑话。你的专业技能相当于你的饭碗，如果你免费将它提供给别人，就会影响你在就业市场上挣钱的能力。现在，通过创建与分享能展示自己才华的数据，你就能建立起个人声望，并传播到你的公司与客户以外的地方。

雇员创建与分享这些可转化为职业声望的数据时，雇主可对此设计出补贴与奖励方案，它超出了销售额与利润等传统的绩效评估指标系统范畴。无论是你的业内声誉，还是你与团队成员或他人进行沟通的效率与速度，甚至由你主持的会议效率目前也可以量化了。例如，当硅谷某家大公司的员工收到公司内部呼叫或聊天请求时，系统会向其显示对方的沟通记录。如果对方给的平均呼叫时间接近半小时，他可决定立即接听来电或将其转入语音信箱。通过这种方式，员工可以更好地管理自己的时间，主管也可以获得有关员工活动的关键性信息。与社交计量标牌所获取的数据一样，该数据让员工用于帮助同事的时间实现了可见性。

该系统与雷伊·达里奥（Ray Dalio）建立的"绝对透明"机制相比不啻"小巫见大巫"。达里奥是世界最大的对冲基金公司桥水联合基金的创始人。桥水基金针对其所做出的决策收集并分析数据，涉及金额高达数十亿美元。几乎所有的会议内容都被记录下来。员工可以记录对同事的

感受，包括挫败感，还可通过苹果平板电脑应用对其他员工的绩效进行评估。所有人都可以查询每个同事的绩效评估情况。该公司的软件致力于发现人们的行为模式，例如通过人们在讨论某个问题时的声音发现其中隐藏的情绪，识别某项决策在何种情况下很少受到质疑，鼓励员工提出质疑并开展内部讨论。许多视频与某些分析均可在员工中共享。达里奥认为，通过将隐性数据变为显性数据，有助于提高认识、改进决策过程并实现更好的成果。我认为，如果桥水联合基金的每名员工都能访问所有的数据并对数据进行修正和实验，那么公司将会取得更优异的成果。

正如我们在约车软件的案例中对司机的讨论一样，通过将职业声望与工作审核相关的数据导出给其他数据服务商，将对1/3的美国人产生重要影响，他们作为独立承包商将会获得一定的收入。某些自由职业项目投标网站与在线协作网站为从业人员提供了创建技能档案与技能合作的平台，并将他们与招募人手完成项目的公司相匹配，例如外包项目平台Freelancer.com网与自由职业平台Upwork网。公司可以评估从业人员的沟通效率、专业技能、工作质量，还能审核从业人员被聘请经历、预算内绩效和规定期限内绩效，以及完工速度方面的情况。在Upwork网站上，你可以导入其他平台的声望数据，包括面向开源及私有软件项目托管平台Github网与技术问答平台Stack Overflow网（它们主要面向软件开发人员），以及设计社区Behance网与创意类作品交流平台Dribbble网（它们主要面向图像设计人员）的数据。与脸谱网上相互验证好友关系相似，评级系统也是对称的，它可以验证雇主与自由职业者是否真正在一起合作过。当从业人员对雇主进行评级时，他们可以指出雇主的工作说明是否准确地反映了项目的范围及所需的时间，还可表明他们是否如期拿到了报酬。过去，只有公司能看到某个项目的所有投标价格，从业人员则需要凭空报价。在Freelancer.com网站与Upwork网站上，所有投标人员都可看到其他投标人员的个人档案、评价与投标情况。这种透明性改变了从业人员与公司间的权利平衡状况。

联合国大会在1948年正式颁布的《世界人权宣言》中指出，人人都有权工作，自由选择职业，享有公正和合适的工作条件。今天，我们需要进一步要求人人有导入、导出个人的记录、评级与评价的权利。社交数据能够给我们提供更公平、透明的方法，在员工自由行使数据权利的前提下，让人尽其才、才尽其用。

在数字课堂上学习

近一个世纪前，美国哲学家约翰·杜威（John Dewey）就指出，“教育不是‘教’与被教的关系，而是一种建设性的动态过程”。但是，教育迄今为止仍然主要建立在权威与讲课的基础上，这种制度的设计旨在为工厂培训工人，打破它需要承担的巨大风险，这就是肯·罗宾逊（Ken Robinson）在其著名的TED演讲《学校扼杀了创造力》中提出的观点。

实际上，2 000年来课堂几乎一成不变。教师站在一群学生面前为其授课，通过考试评估他们所掌握的内容多少。教师只能在期末考试分数批改出来之后，才能知道哪些学生掌握了教学内容，但这已经太晚了。学生很少有机会从自己的同学那里学到知识，也无法将自己课上学到的知识与课外学习的知识相结合。信息通常是单向传输，且所有授课对象听到的是相同的内容。

在过去的一个世纪中，我们越来越多地将注意力放在测试学生对知识的掌握程度上。但事实并非总是如此。苏格拉底给他的学生柏拉图上的每一堂课都采用提问题的形式。对苏格拉底来说，教学生如何提出好问题比为他提供现成的答案更重要。这一观点至今仍具有现实意义，因为搜索引擎能为我们想到的几乎所有问题提供答案（但答案正确与否又是另外一回事）。或者就像哈佛大学物理教授埃里克·马祖尔（Eric Mazur）所说：“你可以忘记事实，但你不能忘记理解。”马祖尔意识到，当他的学生通力合作、共同研究某个问题时，与以往他“一言堂”的授课方式相比，学生们学到了更多的知识，学习能力也变得更强。这促使马祖尔发明在线教育系统——基于云端的学习分析与管理系统Learning Catalytics。

詹妮弗·柯蒂斯（Jennifer Curtis）是首批使用该教育系统的教师之一，她在缅因州的一所高中里教物理课。开始上课时，她会让学生取出苹果平板电脑。她并未将此类设备视为对学习的干扰，而是让学生打开Learning Catalytics系统，在一个小时的时间里在软件指导下学习。首先，柯蒂斯的学生需要根据自己在昨天晚上完成老师布置的作业时阅读的书籍或观看的视频回答一些问题，这让柯蒂斯能根据学生的回答为他们两两配对。然后，柯蒂斯让每一组学生相互说服对方相信自己的回答是正确的，并将他们的最终答案输入该系统。柯蒂斯可以监控到谁更需要帮助。学生熟悉整个过程后，就会学会如何通过辩证的提问、演绎、

团队合作来解决问题。这种方法并没有让学生记忆答案，而是帮助学生理解问题的概念和根本性结构，从而提出各种可能的解决方案。Learning Catalytics系统使柯蒂斯的学生提高了分组讨论的水平，并且表现更好。

与此教育系统的每一次交互都会留下痕迹，这些痕迹可用于提高学生的学习能力。新型教育计划正在陆续推出，它们可以将所有可记录的数据记录下来。例如，这些数据对密涅瓦大学的凯特研究生院（KGI）具有核心意义。该校由喀嚓鱼公司（照片分享与冲印服务公司）的前首席执行官本·内尔森（Ben Nelson）与心理学家、哈佛大学社会科学学院前院长史蒂芬·科斯林（Stephen Kossly）创立。在密涅瓦大学于2015年秋季开学以来，学生与教师在7个城市中开展教学工作，他们主要通过电脑授课。他们发送的视频、聊天信息、对问题的回答都会实时受到分析，以帮助教师调整课堂讨论节奏并成功地让学生加入远程讨论。之后，情绪表现领域专家将对视频编码，以识别出学生出现兴奋、厌烦、挫败、困惑或其他情绪的情况。学生可以获取这些数据，还可在阅读书目和课外参加活动方面获得建议以提高其学业水平。这是一项全面感知的计划，它可以注意到学生开小差的情况，并建议老师下课休息。

通过利用社交数据开展发人深思的实验，将表明学习环境与条件所产生的各种影响。某些学校对一周内每天的课程进行轮换，学生如果能在早晨（或下午）集中注意力，就可能在每门功课上取得优秀成绩。新泽西州的一家高中使用名为Schoology的在线平台，以便对“在家上学”的教学方式开展实验。作为一名科学家，我会系统化地对可更改的各项输入进行各种实验，例如教学风格、教室温度、午餐的伙食或教室里两个好友的课桌之间的距离，并分析这些因素分别对每名学生的学习与成长所产生的影响。教师与学生都能更好地深入了解学生的成长过程，而并非仅靠为数不多的考试。学生与家长都必须有权访问这些结果，并对此进行解读。

为了全面了解并准确评估学生的学习模式，需要先在小学与中学阶段记录数据，然后记录大学以及之后的数据，也许还应包括“终身学习”计划。迄今为止，人们在争论应保存哪些数据和保存多久的问题时，核心都是对隐私权的考虑，这是可以理解的。某个最受欢迎的学生行为跟踪应用选择在学年结束时删除这些数据，这是令人惋惜的损失！教育历史数据不应被删除，而应被保留下来，供学生、家长、教师、教育决

策者能不断地从中了解情况。将各个应用的数据导出到某个综合性教育数据服务商那里，有助于对学生的学习方式进行更深入的了解，数据服务商可以由教育部门、私人基金会或营利性公司创立。同时，还要确保家长、教师、学校管理人员监控数据服务商的数据安全、隐私效率、数据回报。

我曾提到导入、导出数据的可能性，因为我估计很多人都有兴趣在教育环境中发现数据怎样预测人们是否适合某份工作或某个职位，就像雇主在筛选简历时将能力考试得分作为一项考量指标一样。与考试分数不同，对学生收集的数据可能表明他们适合某些类型的工作。因为根据儿童对挑战性场景的反应，可以预测出他们今后的人生，包括他们在求职受拒时的反应。

在预测意料之外的长期结果时，我常使用的一个例子是心理学家沃尔特·米歇尔（Walter Mischel）的“棉花糖实验”，这是他在20世纪60年代在斯坦福大学宾格幼儿园对4~6岁的儿童所做的实验。米歇尔让孩子们自己做出选择（小孩子都喜欢棉花糖）。米歇尔将棉花糖放在桌上，并告诉他们现在就可以从桌上拿一块棉花糖吃，但如果能忍耐15分钟，就可以得到两块棉花糖。说完，米歇尔就离开了房间。他发现有1/3的孩子抵制住了诱惑，而其他孩子则无法控制自己吃掉棉花糖的冲动。一般来说，年龄越大的孩子越能控制自己的欲望。更有趣的是，在这些孩子长大后，米歇尔又与他们进行了联系。他发现当初能忍耐15分钟的孩子比同龄人的能力更强，成绩也更好。学前阶段的意志力与今后的人生是否成功之间的紧密联系令人意外，而且这种关系在之后的研究中也经受住了考验。

教育是一系列棉花糖实验。学生在数字课堂中学习时，可以从更加个人化的反馈中获益，在他们毕业时还可以获取能反映其意志力与其他性格倾向的数据。人们似乎很希望将这些数据用于预测他们今后的人生。雇主希望对应聘员工的性格进行评估，许多公司都要求应聘者在面试中接受心理测试。未来，人们在求职时不仅需要提供简历，还可将自己的教育数据导出提供给未来的雇主，以便向其展示自己的意志力或其他品质。但是，你可能首先要将数据导出给通过数据分析为你提供建议的教育顾问或职业顾问，以便了解自己是否适合从事某项教育或职业。

教育数据应当为人们提供机会而非减少机会，整个教育体系都将通过收集和分析社交数据获益。由于资源有限，教师经常不得不将自己的

主要注意力放在某些学生的身上。他们应当将重点放在分布曲线左侧低于平均成绩的学生、分布曲线右侧高于平均成绩的学生，还是放在分布曲线中间的学生呢？我父亲在教高中理科课程时，对低于平均成绩的学生十分重视，避免他们中途辍学。我在大学教书时，优等生对我的教学最满意，因为我能激发他们的灵感，让他们创意迸发。某些教师则将重点放在提高中等生的成绩上，因为他们通常占到班级学生的大多数。

我们很难对需要重视的群体做出选择，学生的数据（有关学生对教学风格、学习条件、挑战等因素的反应）也不会告诉我们全社会应当重视哪些因素。这个问题并非只有一个答案。全社会必须在学习的等式中先规定好条件，再输入数据并调整权重，之后通过反复实验寻找不断改进的方法，还要区分不同的学生与班级。

精确地界定我们对数据的需求

> 随着医学的数字化进程与障碍的扫除，我们从摆弄翻盖式手机转变为颠覆整个医疗模式。
>
> ——埃里克·托普尔博士（Dr. Eric Topol）

1895年首次发现X线时，这项新发明还很吓人。当时，有一位妇女接受了手部X线照射。据报道称，她看到拍摄出的照片后说了一番令人毛骨悚然的话，“我看到了我死后的样子”。她就是X线的发现者、首位诺贝尔物理学奖得主威廉·伦琴（Wilhelm　Röntgen）的妻子。这项发明使医生立刻就能看到人体内部的情况，从而诊断疾病，提出治疗方案，更精准地进行手术治疗。6个月之后，X线机就为两名在埃塞俄比亚作战负伤的意大利士兵进行检查，以探测子弹在他们体内的位置。

在过去一个世纪中，医学实践已通过一系列科技的发明经历了巨变，这使我们能够以新的方式观察人体。20世纪70年代，人们利用X线进行计算机断层成像，即CAT扫描，使计算机可根据多张X线片对人体建立三维视图。核磁共振成像（MRI）可以检测人体中水与脂肪的含量，我们不仅能看到骨骼，还能看到软组织和血流。2004年，国际人类基因序列协会公布了完整的人类基因序列，这使得精准医疗成为可能，医生可通过分析某人的基因找出最好的治疗方式。如今，全世界已有数

千万台X线机、数十万台CT机（计算机X线断层摄影机）与核磁共振成像仪、数千台基因测序仪。

但是，这些医疗设备的数量与人们日常使用的医疗科技产品相比不啻小巫见大巫，它们就是10亿部智能手机与新兴公司Fitbit、佳明公司、Jawbone公司、智能手表厂商Pebble等公司生产的1亿部生理活动追踪器。通过持续记录人们的关键性生理迹象、锻炼习惯、睡眠模式及情绪方面的情况，这些装置收集的数据对于监测并管理人类的健康与幸福具有核心作用。活动追踪器甚至可以显示出你多久与伴侣（或他人）发生性关系等私密信息。我们从未对我们的身体和自我有如此丰富、翔实的认知。我认为，忽视这些数据的潜力不亚于玩忽职守。但是，在使用它们时需要重新思考病人的权利和医学的数据量。

人们通常只会在患病后去看全科大夫，或每年参加一次体检。在就诊过程中，医生会进行各种测量，例如测量心率、血压、体重，但这通常无法了解你的整体健康状况。如果你存在严重的健康问题，你可能就要接受更多的检查或去看专科大夫。埃里克·托普尔博士称，每年的定期体检效率极为低下，因为在常规检查中，医生只能收集到很少的数据，难以发现问题，但却耗费了大量的时间。缺少及时的信息可能导致许多健康问题的发生。在绝大多数情况下，你的医疗记录中主要是零散的检查结果、明确的症状与主要用于结账的医疗代码。健康数据的收集与使用主要用于满足医疗机构的需要，例如诊所、药房、医院、医疗保险商，数据收集与分析的重点必须转变为满足患者的需要。

这不仅需要改变支付体系的构建方式，因为许多患者不愿分享关于自身健康的信息，甚至对他们的医生也不愿透露。他们可能害怕被诊断出饮食习惯或其他习惯对健康有害，也可能希望问题自行消失。他们还可能认为，诊所和医疗保险公司都与其他公司一样，无法妥善保护客户的数据。有些人已经发现自己的医疗数据对自己造成了不利，例如他们可能因为自身的某种健康情况导致保费提高，甚至被拒保。如果我们能消除人们对自己的医疗记录可能造成不利的担心，这些数据就可以在长期和短期都极大地改善人们的健康与福祉。

另一个问题与此密切相关。尽管患者的所有医疗记录对其人生的某些重大决定十分关键，但直到最近，向患者提供这些医疗数据才成为标准的做法，其中包括检查结果。这是因为医生在写医嘱时并没有站在患者的角度考虑问题。正如某位知名的内科医生对我所说，医生写医嘱的

原因主要有三个：（1）对自己的服务开具详细的账单；（2）回顾患者的病情，就像记住了患者的情况一样；（3）为防止出现问题保留记录，以便在法律诉讼中为自己提供辩护证据。这些因素中很少与改善患者的健康状况有关。

哈佛大学医学院的汤姆·德尔班科博士（Dr. Tom Delbanco）推出了一项名为"公开病历"（Open Notes）计划，在为患者提供医嘱方面取得了飞速的发展。该计划从2010年刚推出时的1.9万名患者用户，发展到2016年的800万名患者用户。通过此项示范性研究，患者可以安全地获取自己的官方医疗记录，其中包含检查结果、诊断结果、处方及医生的嘱咐与建议。无论何时，只要医生添加医嘱，该系统都能通过电子邮件通知相应的患者。每5名患者中就有4名阅读了医嘱，而且他们称这种透明性让他们更好地了解到自己的健康情况，从而改善了医患关系。此外，参加计划的患者还享有修正数据的权利，可以记录处方的副作用或指出医生误解或误诊的问题。例如在治疗的过程中，医生的医疗记录中称患者平均每晚喝5杯酒，而实际情况是患者每周喝5杯酒。随着这项计划的拓展，越来越多的患者开始申请对自己的医疗信息进行修改，并请求医生对医疗术语做出解释。

"公开病历"计划还解决了一个问题，即让心理疾病的患者能看到心理治疗师与心理健康咨询师在治疗过程中的医嘱。心理治疗师通常认为，如果让患者看到自己心理健康方面的医嘱，会对他们今后的人生与幸福造成伤害。这种过度保护、家长式的心理疾病疗法，与20世纪中叶及此前所采用的标准做法并无不同。不允许女性了解自己的医疗信息，因为她们不够坚强。医生只向她们的丈夫或父亲说明病情，交流医疗方案。德尔班科解释称，他认为"心理疾病的患者与膝盖受伤的患者一样，都有权也应当看到医嘱"。此外，通过在医疗决策方面为患者提供更大的透明性与主动性，减少了患者的孤立感与焦虑，提高了他们对医生的信任感，促使他们更积极地改变自己的行为，从而改进长期治疗结果。患者可以审阅自己治疗过程中的医嘱，并提醒自己曾与心理治疗师讨论的应对机制与其他工具。

通过修正医疗记录，还可以以其他方式改善对患者的治疗。医生可以建议患者在一家药房购买所有处方药，这样计算机就能完成它最擅长的工作——记住处方中哪些药物已被购买、何时购买，并标记出不同医生所开处方中的药物一起服用时可能产生的不良反应。但确保这项关键

性服务的提供，不能仅指望患者在同一家药房购买所有处方药，而应将所有医生的医嘱与处方汇总为一份患者医疗记录，这样更安全。你还可以从自己的生理活动追踪器中自动上传数据，以修正自己的医疗记录，还可以利用根据食物照片估算卡路里的实验应用程序Im2Calories等应用修正医疗记录。该应用由谷歌公司研究科学家凯文·墨菲（Kevin Murphy）的团队开发，它能将用户所吃食物的照片转化为食物日记和卡路里计算结果。之后，你就能收到警示信息，例如在你准备享用一瓶葡萄酒时，因为这可能导致你服用的药物产生不良反应。

医疗服务提供商正努力说服患者分享新的数据，这是其计划的组成部分。探索健康公司是南非一家保险公司，它推出了名为“活力”的促销计划。通过与超市、体育用品零售商、医疗用品商店合作，为购物者提供金钱奖励。如果客户的购物卡或信用卡购买记录表明他们购买了健康食物，就能对他们的零售商品账单提供返现或保费折扣激励。如果客户能获得保费降低的条件，可能愿意分享更多的数据。你可能会同意每周多走一些路，并通过手机上的地理位置数据加以验证。这些数据需要通过指纹、直播视频，或其他特殊的识别方式进行验证，类似于支付验证系统。其目的旨在向保险公司证明携带着你的手机行走的是你本人，以防止你让别人拿着你的手机替你行走。

2012年，我与社交数据实验室和联合健康保险公司在斯坦福大学成立了工作小组，研究通过传感器数据改善患者医疗结果的其他方式。在其一种假设情况下，我们思考当患者选择独自居住而非搬入陪护式老年公寓时，如何做到既能对他的状况实施监控，又不必让护士反复到他家进行不必要的探访，因为这样做的成本很高。患者卧室地毯的上方或下方安装了传感器，可以探测到他在起床时是否摔倒在地；安装在他手机上的传感器，可以捕捉到他在其他房间中摔倒的声音。网络摄像头可以拍下视频，以便用情绪识别软件对他的情绪进行分析。如果分析表明他需要帮助时，就可以请他的邻居或医疗专业人士到他的家中陪他。这些科技都不需要有人应答或按按钮，这就是当前大多数医疗报警系统的工作原理。

我个人希望能够对我母亲和她的健康状况进行监控。我母亲已经90多岁了，她住在德国弗莱堡的陪护式公寓中，这样我就不用频繁地去查看她的状况。我提出想在她的公寓中安装网络摄像头，她十分赞同，但却并未完全理解怎样对视频进行分析以发现问题，因为我并不打算每周

7天、每天24个小时不间断地观看视频。但当我着手做此事时，却得知安装摄像头触犯了护士、帮工及其他工作人员的隐私权。我想问的是："他们有什么不可告人的行为不想让我知道？他们难道不是在完成向我们收费的服务项目吗？"如果老年公寓的老板能将视频提供给老人的家人，同时模糊处理我母亲房间内的工作人员的面部，效果就会更好。这将解决所有人的后顾之忧，最重要的是，我希望我母亲能够真正得到优质的照顾。

导入、导出数据的权利还适用于医疗领域其他方面的改善。你可以将医生给你开的药品相关的数据导出给数据服务商，以迅速发现哪个医疗保险计划与参保药房给你的总报价最低。你还可以将数据导出到类似于条件触发网站IFTTT（"让你的网络行为能够引发连锁反应"）的服务公司，它允许你自行设置参数，以便在数据符合要求时触发行动。例如，如果某天的紫外线指数很高，你就会收到应当涂防晒霜的提醒。你还可以将家中的电子装置与个人目标建立联系，例如，在你的传感器表明你当天已经走了10 000步之后，你的电视机才可以打开。

未来的医疗发展趋势是根据每个人不同的基因为其提供量身定制的饮食、药物与其他治疗手段。基因的差异决定了疾病在你体内的发作方式，以及你的身体对病毒、细菌、化学物质所做出的反应。医生将根据你的基因给你开药方，因为基因会影响不同药物成分对人体产生的作用。但是，基因测试还能透露更多信息。由于你的基因与你亲属的基因极为相似，了解你的基因构成就意味着了解你父母、兄弟姐妹、子女的基因构成。如果你分享了自己的数据，可能导致你亲人的数据也被泄露。萨曼莎·克拉克（Samantha Clark）是第二位将自己的基因数据上传到开源的基因研究库openSNP中的人，她在上传基因数据之前曾与自己的家人讨论了这一决定，因为她的基因数据会透露出家人的相关信息。在此领域中，较高的数据安全与隐私效率评级极其关键。无论出台什么法律，你绝不能因自己不应负责或无法改变的事情而受到处罚，基因就是其中一个例子，这同样适用于你的亲人。你应对你的基因数据保留完整的访问记录，它将帮助你预测数据可能对你造成不利的情况。

个人数据在医疗领域中的使用程度还不够。但就像尼古拉斯·克里斯塔基斯（Nicholas Christakis）与詹姆斯·富勒（James Fowler）在他们合著的专著《大连接》中指出的，我们的社交图片与社交生活对我们的健康也产生了巨大的影响。这不仅是重大的公共问题，例如通过谷歌病

毒趋势跟踪流感病毒的传播，也不仅是拥有大量朋友减少自杀的可能性现象（他受朋友自杀影响的情况除外）。克里斯塔基斯与富勒发现，某些健康状况并非人们传统上认为由个人行为或基因所致，而是因为社交行为，例如肥胖。如果某人（在情感和生理上）和你很相似，他常常吃大餐，那么你也容易去吃大餐。因为朋友的体重增加了，他不会在你的体重增加时对你做出负面评价，所以你很可能不担心自己长胖。

两位作者的研究充分表明，用传感器测量人们的体重或血压远远不够，人们还需要注意社交图谱。如果对社交网络开展实验以发现风险因素，并观察这对你的长期健康状况所产生的潜在影响时，会令你大开眼界。你甚至能通过人脸识别软件，对平均体重变化的监测探索与某个圈子的好友来往对你产生的影响。



图7–1 看到你身体内部的情况：1895年的第一张X线片，这是伦琴太太的手
资料来源：照片来自美国国立卫生研究院医学图书馆。

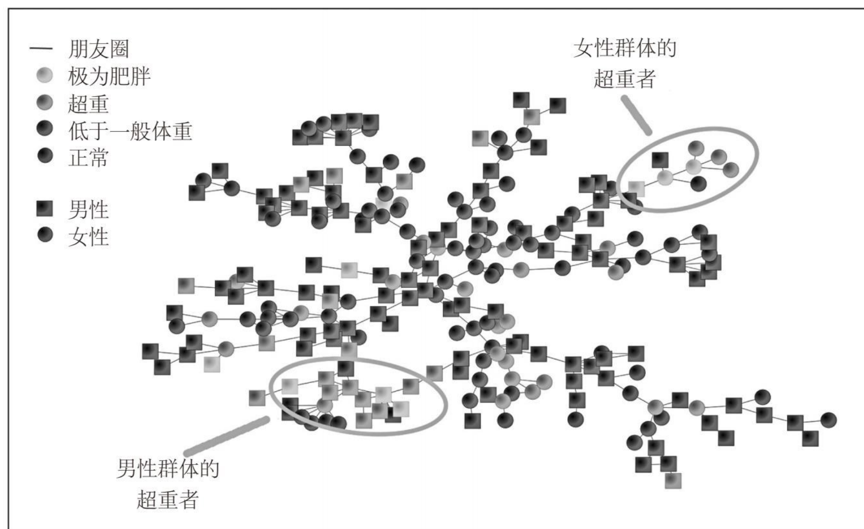


图7–2 看到你身体内部的情况：2015年脸谱网上的一群肥胖的好友
资料来源：照片来自詹姆斯·富勒与尼古拉斯·克里斯塔基斯的专著《大连接》。

美国政府认识到精准医疗的重要性，并通盘考虑了基因学、生活方式与环境等因素对人类健康所产生的影响。美国政府对这项雄心勃勃的研究提供资金，有100万名志愿者愿意向该研究项目分享自己的大量数据。该项目旨在对心脏病、糖尿病、肥胖症、抑郁症等疾病的发病因素展开研究，同时寻找有利于人们健康的选择。随着越来越多的数据服务商进入该领域，医疗的本质将不可避免地发生变化。医疗将由对行为的测量转变为对行为的改变，由被动的治疗转变为预防性医疗，由描述性医疗转为预测性医疗。患者能更好地观察自己当前的行动将会如何影响其未来的健康状况。这对于有关饮食、锻炼、睡眠方面的决定尤为重要，因为反馈需要多年时间，而且各种因素的影响相互交织。医生应当让患者参与到探索不同假设情况的过程中来，鼓励患者积极地获取、修正医疗数据，对医疗数据展开实验，将其导出到不同的数据服务商那里，并参与对本人医疗记录的管理当中。其结果将使人们拥有比医疗数据更重要的东西——健康。

随着人们越来越习惯于主动管理自己的健康，越来越多的人将把自己的幸福当作自己的责任而非医生的责任。针对通过改变人们行为就能预防的健康问题，我们能减少此类问题的产生吗？人们不再因为之前的某种健康状况而被保险公司拒保，而是要考虑自己“选择某种生活方式后”可能会被拒保或多缴保费。有时，数据需要我们做的不亚于苦口良药。

决策的量化

假设我们能够获得数据透明性与主动性方面的权利。我们能够访问有关自己的所有数据，包括与他人共同创建的数据和共同拥有的数据。我们可以根据数据安全、隐私效率、数据回报等指标对数据服务商进行检查和对比。我们可以修正并模糊处理我们的数据，精确地了解通过分享更多的数据所能获得的好处与需要承担的风险。我们可以对数据服务商展开实验并导入、导出我们的数据，以便更好地了解它们能为我们做什么。

开展实验比较容易；我们可以调整各种输入与算法，并观察其对产出会产生什么影响。如果我们知道想要优化的对象，优化也会比较简单。但是，这并不意味着不会出现难以解决的问题。

我们面临的一大难题是如何定义公平。如果我们只能利用极为有限的数据分配稀缺的资源时，我们就会尽最大的努力确保公平。我们会让人们在市场中对资源展开竞价；如果市场不是合理的分配机制，我们会制定诸如“先到者先得”等规则或采取随机摇号方式。但是，随机摇号并非我们唯一的选择。当患者需要做器官移植手术时，我们不能通过抛硬币的方式决定由谁获得唯一的肝源。我们可以将这块肝脏送到拍卖行，价高者得。我们也可以让医生来做出这个攸关患者生死的艰难决定，医生必须依据患者健康方面的数据确定优先次序。不幸的是，某些人无法很快接受手术。2002年，诺贝尔奖得主埃尔文·罗斯设计出一套创新算法，以优化器官捐赠者与患者之间的配对。我们未来会利用社交数据概算出每名患者生命的价值吗？我们能极为精确地预测出患者每多活一年，对其家人乃至全社会产生的价值，并将其作为我们算法的输入项之一吗？我们会将对寿命数量的测量转变为对寿命质量的测量吗？对寿命质量的界定更复杂、更具争议性。即便每个人都同意确定优先顺序时所考虑的各项参数，我们仍需要在这个性命攸关的公式中确定每一个参数的权重。

这不仅适用于上文中深层次的哲学问题（由谁来接受捐赠器官以获得生存的机会），而且适用于更加实际的问题，例如如何将车停入公共停车场。传感器可以对空余停车位提供极为精确的数据。博世公司的自动驻车上坡辅助系统利用摄像头测量停车位的大小，并指导车辆进入停车位。摄像头在车辆行进中定时捕捉图像，这些图像与全球定位系统结

合后，就能实时识别空余停车位的位置。许多汽车制造商目前都提供了自动驻车系统，这使得汽车制造商能形成广泛覆盖的网络。这有点儿像自动化公司Vigilant行车记录仪组成的网络，它可以收集整个国家的车辆牌照信息。

如果司机将自己的目的地分享给连到博世公司数据库的导航应用时，就能获得提醒，得知车辆即将经过的公共停车场是距离目的地最近的停车场。此外，这款应用甚至能给出前往最近的空余停车位的路线，而非前往目的地的路线。该系统给司机与城市都带来了好处，因为它减少了交通压力和尾气排放。

这种方法比现有的系统更先进，因为你现在只能先行驶到目的地，再通过碰运气的方式找停车位，但它仍然存在很大的偶然性。为减少不确定性，某些开发人员正在考虑如何以合理的条件鼓励人们分享免费停车位的相关信息。猴子停车是一款停车应用软件，它允许用户设置转让停车位的价格。该应用提高了透明性与主动性，但人们却认为这有失公平。旧金山市不允许在本市范围内使用这款应用，并称这会造成危险驾驶，因为司机在驾驶过程中需要编写短信；而且违反了法律。阳光明媚的圣莫尼卡市也不欢迎这款软件，负责该市停车事务的主管称："他们并不是停车位的所有者。这既违法也不道德。这相当于街头混混站在一块空地前，招手让人们在这里停车，然后向对方索要小费。"因此，这款应用软件转变了策略，只要私人的路边空车位和车库空位可用于租赁，且价格与其他停车位相比具有竞争力，就可将司机与之配对。

对猴子停车应用软件的批评意见，主要集中在它将使公共停车位的价格飞涨，这让许多人无法承受。但是，我们来看看市场机制发挥作用的情况。假设某人名叫多特，她是圣莫尼卡市皮科大道海滩上的百叶窗海滩酒店的前台，她听到自己的手机提示音后得知："啊！如果我现在把我的车从车位上移走，有人愿意为这个车位付给我40美元。"此时在街道上，一辆迈锐宝轿车正在街区周围转悠，司机正在焦急地寻找圣莫尼卡市法院附近的停车位，因为今天是法庭开庭日。多特接受了他的开价，请了一会儿短假去把自己的车开走。如果她一时半会儿找不到车位，就会受到主管的警告，或者不得不去付费停车。如果她能很快找到免费车位，她就能赚到40美元。我们不清楚这款应用软件是帮助还是伤害了愿意用空闲时间换钱的人，他们的时间同样有限。

但是，公共停车位市场可能会被一小撮人主宰和控制。精明的企业

家可能会买下许多辆廉价轿车，并雇用司机在每天很早的时候去占据许多利用率最高的停车位。这可能使多特这样的上班族很难找到停车位，更不要说把自己的车位有偿租给别人了。无论是免费停车位还是付费停车位，人们将越来越难以按正常的价格得到车位。为解决停车位不足的情况，市政部门可能会对猴子停车软件等车位交易市场中的商家征税，以利用这笔资金改善公共交通。此时，数据为所有公众而非一部分人服务。

通过各种方式，社交数据的革命使之前从未量化或无法量化的一切事物都能被量化。过去，我们有理由说，我们无法利用数据或工具对全社会所面临的选择进行归纳和分析。这种情况现在已经一去不复返了，我们可以让自己的选择更加个性化，并观察由此产生的影响。当然，这并不容易实现。

透明性与主动性将推动我们向具体目标迈进，但它们并没有为我们确定目标。此外，没有一种“放之四海皆准的设置”能为所有人优化数据所用；即便我们能够完美地进行一切测算，每个人对各项权重的分配也不一样。未来，我们可能会通过分析一系列数据，极为精确地预测人们的健康与幸福情况，并据此对各项选择排序，这些数据包括人们的搜索条件、社交图片、基因、脸部表情。如果你根据自己在大学所学的知识及之后打算选择的职业道路了解到自己患心脏病的风险很大，你会做出不同的选择吗？你会更换工作、医疗保险或居住的城市吗？你在访问数据、检查数据、修正数据、模糊处理数据、对数据开展实验、导入和导出个人数据时，就能更好地了解自己的目标、关注点，并对自己个人健康函数中的各个变量设定权重。通过体验你在考虑不同假设情况时的感受，你将会坚持自己的价值观，在必要时还会调整自己的公式。

我们现在有能力对艰难决定中的取舍进行量化，突出我们的价值观，并测算由此产生的结果，这促使我们在公平与不公平之间做出选择，我们再也不能选择视之不见，也不能选择碰运气。当我们有能力对世界上一切事物的数据进行挖掘，在透明性与主动性方面行使我们的权利时，我们的数据将服务于我们。

---

1. 1英亩≈0.004平方千米。——编者注



走出洞穴，沐浴阳光

如果他们不能转头，那么他们如何看到影子之外的事物呢？

——柏拉图

在伯罗奔尼撒战争中，苏格拉底有时会与自己的弟子、柏拉图的哥哥格劳孔坐而论道，并向他展示知识的来源。苏格拉底认为，通过让光照进现实世界，真理就能显现出来。在“洞穴之喻”中，一群人被迫一生都居住在某个黑暗的洞穴中。他们的脖子被锁链绑住，无法转头。在他们面前有一面墙壁，在他们后面生起一堆篝火，有人拿着各种假人假物，在火堆前前后走动。火光非常强烈，足以使这些从小就被囚禁在洞穴内的人看到映照在眼前洞壁上的人影活动——前后移动、相互交流、交换物品。有时，这些影子从一个地方突然跳跃到另一个地方，有时它们的动作具有连贯性，并且井然有序。洞穴中的人们日复一日地观看这些人影，并将外界的声音与这些人影联系在一起。他们对世界的认识仅限于他们看到的这一切。

之后，苏格拉底假设洞穴中有个人有机会转动自己的头。由于长期身处黑暗，他的眼睛已经适应了黑暗，能看到极其昏暗的影子。当他朝篝火看去时，强烈的光线令他暂时失明。他努力想看清，但却很困惑、很沮丧，于是他又转回头面向洞穴中的黑暗，此时他又能看见了。他甚至可能会告诉其他人，除了洞壁上的影子他看不到任何东西。

之后，这个人获得了自由。火光再次令他暂时失明，但这次他有充分的时间适应它。最终，他自由了！不久后，他发现其实是假人在前后移动，并投影在洞壁上——他认识到假人与这些影子的关系。

最后，这个人离开了洞穴，沐浴在阳光下。此时，他明白自己需要时间适应光线，他很有耐心。他看到了影子，但他认识到这些影子并非现实。

如果他返回洞穴，努力劝说其他人走出洞穴、沐浴阳光，他可能会遭到拒绝。回到黑暗的洞穴之后，他会再次暂时失明，无法看清假人的影子。由于他再也无法看到任何东西，洞穴中的其他人可能认为这个人的眼睛已经被光线毁掉了。谁会责备这些人呢？毕竟没有人愿意自己的世界被颠覆。

柏拉图在2 000多年前讲述了苏格拉底的这次对话。如今，我们面临的情况与之极为相似。脸谱网与谷歌等数据服务商推出的项目就像在洞壁上投射的影子，它可以让我们解读。与“洞穴之喻”中的影子一样，我们生活中留下的电子踪迹也是真实世界的产物：谷歌不会编造网页作为我们的搜索结果，脸谱网也不会凭空捏造好友的发帖以骗取脸谱网新闻馈送的用户。同时，我们通过数据服务商带来光明，它们帮助我们理解海量数据的创建方式。如此大规模的交互与活动不仅令柏拉图无法想象，我们也会深感震惊。

但是，黑暗中发生的事情很多。如果算法投影在洞壁上的所有内容都无法转化为客观现实，我们就会遇到风险。我们需要时间来适应这些新型的数据来源，并了解如何利用工具帮助我们观看和使用这些数据来源，甚至享受它们。透明性相关的权利将使我们看到光线的形态并了解影子的形成过程，且不会出现暂时失明；主动性相关的权利将使我们能根据自己的需要，改变并移动光源。

黑暗的洞穴生活已经一去不复返了。与柏拉图的囚徒不同，我们的头并未被锁链束缚。即便付出大量的工作，我们也必须自由地观看，自由地行动。

即便一开始光线极为耀眼，我们也必须这样做。



我要对我在伦敦的撰写人罗宾·丹尼斯（Robin Dennis）深表谢意，他将我在讯佳普上的100次演讲和1 000个小时的通话内容编入本书中。我要感谢我的文稿代理人吉姆·莱文（Jim Levine）。我还要感谢基础读物出版社的编辑T·J·凯莱赫（T. J. Kelleher），没有他，本书就无法面世。

还有许多人需要感谢。根据本书的宗旨，我邀请你们行使自己的权利，修正我在weigend.com/thanks上创建的数据，这是对本书撰写提供帮助的所有人名单，包括我的朋友、助手、同事、学生。

我还想知道当你要导出数据时，数据服务商会对你说什么。你可以在ourdata.com网上发现更多有关本书的内容，还可在weigend.com网站上获得有关我的更多信息。我的电子邮箱是andreas@weigend.com。

2016年8月

于中国上海和美国旧金山完稿